THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Cotutelle internationale : Université de Valladolid

Présentée et soutenue par

Paula GORDALIZA PASTOR

Le 30 septembre 2020

Fair Learning: une approche basée sur le Transport Optimale

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Mathématiques et Applications**

Unité de recherche :
IMT : Institut de Mathématiques de Toulouse

Thèse dirigée par
Fabrice GAMBOA et Eustasio Del BARRIO

Jury

M. Gabriel PEYRÉ, Rapporteur
M. Gilles BLANCHARD, Rapporteur
M. Gábor LUGOSI, Examinateur
M. Massimiliano PONTIL, Examinateur
Mme Amandine MARREL, Examinatrice
M. Fabrice GAMBOA, Directeur de thèse
M. Jean-Michel LOUBES, Co-directeur de thèse
M. Eustasio DEL BARRIO, Co-directeur de thèse

Abstract

The aim of this thesis is two-fold. On the one hand, optimal transportation methods are studied for statistical inference purposes. On the other hand, the recent problem of fair learning is addressed through the prism of optimal transport theory.

The generalization of applications based on machine learning models in the everyday life and the professional world has been accompanied by concerns about the ethical issues that may arise from the adoption of these technologies. In the first part of the thesis, we motivate the fairness problem by presenting some comprehensive results from the study of the *statistical parity* criterion through the analysis of the *disparate impact* index on the real and well-known *Adult Income* dataset. Importantly, we show that trying to make fair machine learning models may be a particularly challenging task, especially when the training observations contain bias. Then a review of Mathematics for fairness in machine learning is given in a general setting, with some novel contributions in the analysis of the price for fairness in regression and classification. In the latter, we finish this first part by recasting the links between fairness and predictability in terms of probability metrics. We analyze repair methods based on mapping conditional distributions to the Wasserstein barycenter. Finally, we propose a *random repair* which yields a tradeoff between minimal information loss and a certain amount of fairness.

The second part is devoted to the asymptotic theory of the empirical transportation cost. We provide a Central Limit Theorem for the Monge-Kantorovich distance between two empirical distributions with different sizes n and m, $\mathcal{W}_p(P_n, Q_m)$, $p \geq 1$, for observations on $\mathbb{R}$. In the case $p > 1$ our assumptions are sharp in terms of moments and smoothness. We prove results dealing with the choice of centering constants. We provide a consistent estimate of the asymptotic variance which enables to build two sample tests and confidence intervals to certify the similarity between two distributions. These are then used to assess a new criterion of data set fairness in classification. Additionally, we provide a moderate deviation principle for the empirical transportation cost in general dimension. Finally, Wasserstein barycenters and variance-like criterion using Wasserstein distance are used in many problems to analyze the homogeneity of collections of distributions and structural relationships between the observations. We propose the estimation of the quantiles of the empirical process of the Wasserstein's variation using a bootstrap procedure. Then we use these results for statistical inference on a distribution registration model for general deformation functions. The tests are based on the variance of the distributions with respect to their Wasserstein's barycenters for which we prove central limit theorems, including bootstrap versions.

Keywords: Fairness, statistical parity, equality of odds, disparate impact, machine learning, Wasserstein distance, repairing methodology, Central Limit Theorem, moderate deviation principle, Wasserstein variation, goodness of fit.

Résumé

L'objectif de cette thèse est double. D'une part, les méthodes de transport optimal sont étudiées pour l'inférence statistique. D'autre part, le récent problème de l'apprentissage équitable est considéré avec des contributions à travers le prisme de la théorie du transport optimal.

L'utilisation généralisée des applications basées sur les modèles d'apprentissage automatique dans la vie quotidienne et le monde professionnel s'est accompagnée de préoccupations quant aux questions éthiques qui peuvent découler de l'adoption de ces technologies. Dans la première partie de cette thèse, nous motivons le problème de l'équité en présentant quelques résultats statistiques complets en étudiant le critère *statistical parity* par l'analyse de l'indice *disparate impact* sur l'ensemble de données réel *Adult income*. Il est important de noter que nous montrons qu'il peut être particulièrement difficile de créer des modèles d'apprentissage machine équitables, surtout lorsque les observations de formation contiennent des biais. Ensuite, une revue des mathématiques pour l'équité dans l'apprentissage machine est donné dans un cadre général, avec également quelques contributions nouvelles dans l'analyse du prix pour l'équité dans la régression et la classification. Dans cette dernière, nous terminons cette première partie en reformulant les liens entre l'équité et la prévisibilité en termes de mesures de probabilité. Nous analysons les méthodes de réparation basées sur le transport de distributions conditionnelles vers le barycentre de Wasserstein. Enfin, nous proposons le *random repair* qui permet de trouver un compromis entre une perte minimale d'information et un certain degré d'équité.

La deuxième partie est dédiée à la théorie asymptotique du coût de transport empirique. Nous fournissons un Théorème de Limite Centrale pour la distance de Monge-Kantorovich entre deux distributions empiriques de tailles différentes n et m, $\mathcal{W}_p(P_n, Q_m),\ p \geq 1$, avec observations sur $\mathbb{R}$. Dans le cas de $p > 1$, nos hypothèses sont nettes en termes de moments et de régularité. Nous prouvons des résultats portant sur le choix des constantes de centrage. Nous fournissons une estimation consistente de la variance asymptotique qui permet de construire tests à deux échantillons et des intervalles de confiance pour certifier la similarité entre deux distributions. Ceux-ci sont ensuite utilisés pour évaluer un nouveau critère d'équité de l'ensemble des données dans la classification. En outre, nous fournissons un principe de déviations modérées pour le coût de transport empirique dans la dimension générale. Enfin, les barycentres de Wasserstein et le critère de variance en termes de la distance de Wasserstein sont utilisés dans de nombreux problèmes pour analyser l'homogénéité des ensembles de distributions et les relations structurelles entre les observations. Nous proposons l'estimation des quantiles du processus empirique de la variation de Wasserstein en utilisant une procédure *bootstrap*. Ensuite, nous utilisons ces résultats pour l'inférence statistique sur un modèle d'enregistrement de distribution avec des fonctions de déformation générale. Les tests sont basés sur la variance des distributions par rapport à leurs barycentres de Wasserstein pour lesquels nous prouvons les théorèmes de limite centrale, y compris les versions *bootstrap*.

Mots-clé: Équité, *statistical parity*, *equality of odds*, *disparate impact*, apprentissage machine, distance de Wasserstein, méthodologie de réparation, Théorème de Limite Centrale, Principe de déviations modérées, variation de Wasserstein, qualité de l'ajustement.

Resumen

El propósito de esta tesis es doble. Por un lado, se estudian métodos de transporte óptimo destinados a hacer inferencia estadística. Por otro lado, se considera el reciente problema del aprendizaje justo con contribuciones basadas en la teoría del transporte óptimo.

El uso generalizado de aplicaciones basadas en modelos de aprendizaje automático en la vida cotidiana y en el mundo profesional ha traído consigo preocupaciones sobre las cuestiones éticas que surgen de la adopción de estas tecnologías. En la primera parte de la tesis, motivamos el problema de la equidad presentando algunos resultados estadísticos exhaustivos sobre el estudio del criterio *statistical parity* a través del análisis del índice *disparate impact* en el conjunto de datos reales *Adult income*. Mostramos que tratar de hacer modelos justos puede ser una tarea particularmente difícil, especialmente cuando las observaciones de entrenamiento contienen sesgos. A continuación, se hace una revisión de los métodos matemáticos para el aprendizaje justo en un marco general, con contribuciones novedosas en el análisis del precio de la equidad en regresión y clasificación. En este último, concluimos esta primera parte reformulando los vínculos entre la equidad y la previsibilidad en términos de métricas de probabilidad. Analizamos los métodos de reparación basados en el transporte de las distribuciones condicionales hacia el baricentro de Wasserstein. Por último, proponemos el *random repair* que establece un equilibrio entre la pérdida de información y el nivel de equidad.

La segunda parte está dedicada a la teoría asintótica del coste empírico de transporte. Proporcionamos un Teorema Central del Límite para la distancia Monge-Kantorovich entre dos distribuciones empíricas con tamaños n y m, $\mathcal{W}_p(P_n, Q_m)$, $p \geq 1$, y observaciones en $\mathbb{R}$. En el caso $p > 1$ nuestras hipótesis son minimales en términos de momentos y suavidad. Probamos resultados que tratan con la elección de las constantes de centramiento. Proporcionamos una estimación consistente de la varianza asintótica que permite construir tests de dos muestras e intervalos de confianza para certificar la similitud entre dos distribuciones. Éstos se utilizan luego para evaluar un nuevo criterio de equidad en clasificación binaria. Además, proporcionamos un principio de desviaciones moderadas para el coste empírico de transporte en dimensión general. Por último, los baricentros de Wasserstein y el criterio de varianza utilizando la distancia de Wasserstein se emplean en muchos problemas para analizar la homogeneidad de una colección de distribuciones y las relaciones estructurales entre observaciones. Proponemos la estimación de los cuantiles del proceso empírico de la variación de Wasserstein mediante un procedimiento *bootstrap*. A continuación, con estos resultados hacemos inferencia estadística en un modelo de deformación general. Los tests se basan en la varianza de las distribuciones con respecto a su baricentro de Wasserstein, para los que probamos teoremas centrales del límite, incluidas las versiones *bootstrap*.

Palabras clave: Equidad, *statistical parity*, *equality of odds*, *disparate impact*, aprendizaje automático, distancia de Wasserstein, metodología de reparación, teorema central del límite, principio de desviaciones moderadas, variación de Wasserstein, bondad de ajuste.

Acknowledgements

First of all, I would like to express my deep appreciation to my supervisors Prof. F. Gamboa, Prof. J-M. Loubes and Prof. E. del Barrio, to whom I am grateful both for their dedication and continuous support, and for the trust they have placed in my work from the very beginning. This thesis is the result of their guidance, suggestions and encouragement. I could not have imagined having better advisors and mentors for my Ph.D study.

Besides my advisors, I would like to thank Prof. P. Besse, Prof. L. Risser and Prof. H. Léscornel, who I've had the honor of working with, being part of this thesis fruit of these collaborations.

To my colleagues in the Institut de Mathématiques de Toulouse and in the Departamento de Estadística e Investigación Operativa for joining me during this intense and enriching experience.

To my loved ones, for supporting me unconditionally throughout pursuing this Ph.D thesis and my life in general.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation and framework for the fairness problem

Artificial Intelligence technologies are undoubtedly making human life easier over the last years. In particular, machine learning based systems are reaching society at large and in many aspects of the everyday life and the professional world. Powering self-driving cars, accurately recognizing cancer in radiographies, or predicting our interests based upon past behavior, are just a few examples in the wide array of technological applications in which they are showing great promise. Yet with its benefits, machine learning techniques are not absolutely objective since model classifications and predictions rely heavily on potentially biased data. Hence this generalization of predictive algorithms has been accompanied by concerns about the ethical issues that may arise from the adoption of these technologies, not only among the research community but also among the entire population. Thanks to this, there has been a great push for the emergence of multidisciplinary approaches for assessing and removing the presence of bias in machine learning algorithms.

Fair learning is a recently established area of machine learning that studies how to ensure that biases in the data and algorithm inaccuracies do not lead to models that treat individuals unfavorably on the basis of characteristics such as race, gender, disabilities, and sexual or political orientation, just to name the more striking. The purpose of this thesis is presenting a mathematical approach for the fairness problem in machine learning. The application of our theoretical results aims at shedding some light on the maelstrom of techniques or mere heuristics that ceaselessly appear to address these issues. We believe that a robust mathematical ground is crucial in order to guarantee a fair treatment for every subgroup of population, which will contribute to reduce the growing distrust of machine learning systems in the society.

The mathematical framework for fair learning is usually presented in the literature as follows. Consider the probability space $(\Omega, \mathcal{B}, \mathbb{P})$, with $\mathcal{B}$ the Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geq 1$. We will assume in the following that the bias is modeled by the random variable $S \in \mathcal{S}$ that

represents an information about the observations $X \in \mathcal{X} \subset \mathbb{R}^d$, that should not be included in the model for the prediction of the target $Y \in \mathbb{R}^d,\ d \geq 1$. The variable S is referred to as the *protected* or *sensitive attribute*, and it is usually assumed to be observed. Finally, the class of measurable functions $f : (X, S) \mapsto Y$ will be denoted by $\mathcal{F}$ and, particularly, $\mathcal{G}$ will denote the class of binary classifiers.

1.2 Bias and definition of fairness in machine learning

One of the first steps is showing the importance of understanding how bias could be introduced into automatic decisions. From a mathematical point of view, we will describe in chapter 3 two possible models, proposed first in Serrurier et al. [2019], that aim at formalizing this issue. The first model (see Figure 3.1a) corresponds to the case where the data X are subject to the bias nuisance variable S which, in principle, is assumed not to be involved in the learning task, and whose influence in the prediction should be removed. Under this assumption, a fair model requires that the outcome does not depend on the sensitive variable. On the other hand, the second model (see Figure 3.1b) deals with the situation when a biased decision is observed as a result of a fair score which has been biased by the uses giving rise to the target Y. Thus, a fair model in this case will change the prediction in order to make it independent of the protected variable. We observe that the probabilistic notion underlying each model is a different type of independence between distributions. Hence, the choice of this assumption is decisive in the criterion used for fairness. In this sense, we will be looking at the notion of *perfect fairness* as an independence between the protected variable S and the outcome $\widehat{Y} = f(X, S)$, both considering conditionally given (second model) or not (first model) the true value of the target Y. Each approach has motivated two different definitions:

- *Statistical parity* (SP) [Dwork et al., 2012] deals with $\hat{Y} \perp\!\!\!\perp S$
- *Equality of odds* (EO) [Hardt et al., 2016] considers $\hat{Y} \mid Y \perp\!\!\!\perp S$, and is especially well-suited for scenarios where ground truth is available for historical decisions used during the training phase.

Most fairness theory has been developed particularly in the case when $\mathcal{S} = \{0, 1\}$ and S is a sensitive binary variable. In other words, the population is supposed to be possibly divided into two categories, taking the value $S = 0$ for the *minority* (assumed to be the unfavored class), and $S = 1$ for the *default* (and usually favored class). Hence, we will study more deeply this case in the first part of the thesis, starting with chapter 2 which is framed in the binary classification framework. Its purpose is to motivate the problem of fairness in machine learning by presenting some comprehensive statistical results obtained from the study of the *statistical parity* criterion with applications to credit scoring. We specifically consider the real *Adult Income* dataset[1]. It consists in forecasting a binary variable which corresponds to an income lower or higher than 50k$ a year, where the existing unbalance between the income prediction and the *Gender* and *Ethnic origin* sensitive binary variables is clearly noticeable. This decision could be potentially used to evaluate the credit risk of loan applicants, making this dataset particularly popular in the machine learning community. In this framework, bias of a binary classifier $g(X, S) = \hat{Y}$ is frequently quantified with the *disparate impact* (DI):

$$DI(g, X, S) = \frac{\mathbb{P}(g(X, S) = 1 | S = 0)}{\mathbb{P}(g(X, S) = 1 | S = 1)}.$$

[1] `https://archive.ics.uci.edu/ml/datasets/adult`

This index was first introduced as the $4/5^{th}$-rule by the State of California Fair Employment Practice Commission (FEPC) in 1971[2]. Since then, the threshold 0.8 has been chosen in different trials as a legal score to judge whether the discriminations committed by an algorithm are acceptable or not (see Feldman et al. [2015], Mercat-Bruns [2016] or Zafar et al. [2017a]). Yet, this score, as well as many others described in the fair learning literature, are often used without statistical control. In the cases where test procedures or confidence bounds are provided, they are computed using a resampling scheme to get standardized Gaussian confidence intervals under a Gaussian assumption which does not fit the distribution of the observations. In this chapter, we promote the use of confidence intervals to control the risk of false discriminatory assessment. Importantly, we obtain the exact asymptotic distribution of the estimates of different fairness criteria using the classical Delta method approach [Van der Vaart, 1998]. Moreover, we show that some standard approaches, including the removal of the sensitive variable or the use of testing technics appeared as irrelevant when trying to correct the discriminatory behaviour of machine learning algorithms. Finally, we will test two a priori naive solutions consisting either in building a differentiate algorithm for each class of the population or adapting the decision of a single algorithm in a different way for each subpopulation. Only the latter proves helpful in obtaining a fair classification.

Returning to a more general supervised learning context, a review of the main fair learning methodologies proposed in the literature over the last years will be presented from a mathematical point of view in chapter 3. Moreover, following our independence-based approach, we will consider how to build fair algorithms and the consequences on the degradation of their performance compared to the possibly unfair case. This corresponds to the price for fairness. Recall that the performance of an algorithm is measured through its risk defined by

$$R(f) = \mathbb{E}(\ell(Y, f(X, S))),$$

with $\ell : (Y, \hat{Y}) \mapsto \ell(Y, \hat{Y}) \in \mathbb{R}^+$ a certain loss function. Theoretically, a fair model $f \in \mathcal{F}$ can be achieved by restricting the risk minimization to a fair class of models, namely, $\inf_{f \in \mathcal{F}_{\text{Fair}}} R(f)$. This class $\mathcal{F}_{\text{Fair}}$ will be particularly denoted by

$$\mathcal{F}_{SP} := \{f(X, S) \in \mathcal{F} \quad \text{s.t} \quad \hat{Y} \perp\!\!\!\perp S\} \qquad \text{or} \qquad \mathcal{F}_{EO} := \{f(X, S) \in \mathcal{F} \quad \text{s.t} \quad \hat{Y} | Y \perp\!\!\!\perp S\},$$

depending on the fairness notion considered. In general, the price for fairness is then computed as

$$\mathcal{E}_{\text{Fair}}(\mathcal{F}) := \inf_{f \in \mathcal{F}_{\text{Fair}}} R(f) - \inf_{f \in \mathcal{F}} R(f),$$

where the $\inf_{f \in \mathcal{F}} R(f)$ is known as the Bayes Risk. This minimal excess risk will be studied in this review chapter, under both fairness assumptions and in two different frameworks: regression and classification. On the one hand, some existing results on the boundness of the price for fairness as *statistical parity* will be recasted. We make the following main points: (i) in the regression problem, we recall a result from Le Gouic and Loubes [2020] giving a lower bound for the minimal excess risk in terms of the quadratic Wasserstein distance; (ii) in the classification problem, we anticipate the upper bound for the minimal excess risk in terms of the Wasserstein variation proposed in paper Gordaliza et al. [2019], which we will refer to later in this introduction since it corresponds to the content of chapter 4. On the other hand, the price for fairness as *equality of odds* is also studied. Importantly, novel results giving the expressions of the optimal fair classifier and the optimal fair predictor (under a linear regression gaussian model) will be presented.

[2] `https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml`

1.3 Imposing fairness with a repair methodology

The importance of ensuring fairness in algorithmic outcomes has raised the need for designing procedures to remove the potential presence of bias. From a procedural viewpoint, methods for imposing fairness can be roughly divided into three families. Methods in the first family consist in pre-processing the data or in extracting representations that do not contain undesired biases, which can then be used as input to a standard machine learning model. Methods in the second family, also referred to as in-processing, aim at enforcing a model to produce fair outputs through imposing fairness constraints into the learning mechanism. Methods in the third family consist in post-processing the outputs of a model in order to make them fair. Yet building perfect fair models may lead to poor accuracy: changing the world into a fair one with positive action might decrease the efficiency defined as its similarity to the uses monitored through the test sample. While in some fields of application it is desirable to ensure the highest possible level of fairness; in others, including Health Care or Criminal Justice, performance should not be decreased since the decisions would have serious implications for individuals and society. Hence, it is of great interest to set a trade-off between fairness and accuracy, resulting in a relaxation of the notion of fairness that is frequently presented in the literature as *almost* or *approximate fairness*. To this aim, most methods approximate fairness desiderata through requirements on the lower order moments or other functions of distributions corresponding to different sensitive attributes.

In particular, in chapter 4 we present our repair methodology, which is included in the first category of methods. There, the notion of fairness through the prism of *statistical parity* is considered in the binary classification setting. Our repairing proposal consists in changing the original distribution of the input variable X conditionally given the protected group S, denoted by $\mu_s := \mathcal{L}(X|S=s), s \in \{0,1\}$, in order to make them equal (*total repair* for perfect fairness) or close enough (*partial repair* for almost fairness) to a new unknown target distribution. More precisely, *total repair* amounts to mapping the original variable X into a new variable $\tilde{X} = T_S(X)$ such that conditional distributions with respect to S are the same, namely,

$$\mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right).$$

Note that the transformation $T_S : \mathbb{R}^d \to \mathbb{R}^d$ is random since it depends on the value of the protected variable S. In this case, any classifier g built with such information will be such that $\mathcal{L}\left(g(\tilde{X}) \mid S=0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S=1\right)$, guaranteeing full fairness of the classification rule.

The Wasserstein (a.k.a Monge-Kantorovich) distance appears as an appropriate tool for comparing probability distributions and arises naturally in optimal transport theory (we refer to Villani [2009] for a detailed description). For P and Q two probability measures on $\mathbb{R}^d$, the squared Wasserstein distance between P and Q is defined as

$$\mathcal{W}_2^2(P,Q) := \min_{\pi \in \Pi(P,Q)} \int \|x-y\|^2 d\pi(x,y)$$

where $\Pi(P,Q)$ the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q. The crucial fact that it respects the structure in the data makes it a good choice for the repairing procedures, as it will preserve the relationship between the outcome and the data. Thus, this choice suggests that the distribution of the repair should be the Wasserstein barycenter μ_B between the conditional distributions μ_s with respect to the weights $\pi_s = \mathbb{P}(S=s)$ of the protected classes $s \in \{0,1\}$, namely

$$\mu_B \in argmin_{\nu \in \mathcal{P}_2} \left\{\pi_0 \mathcal{W}_2^2(\mu_0,\nu) + \pi_1 \mathcal{W}_2^2(\mu_1,\nu)\right\},$$

and that the optimal way to reach it are the optimal transport maps $\mu_B = \mu_s \circ {T_s}^{-1}$, for $s = 0, 1$. Note from the definition of the Wasserstein barycenter that this repair methodology could be easily extended to S multiclass.

As mentioned before, we justify such an approach providing in Theorem 4.3.3 an upper bound for the price for fairness of the transportation towards the barycenter μ_B. More precisely, we prove that the minimal excess risk when considering the best classifier g_B (Bayes rule) with the repaired data and the original data is upper bounded by the weighted Wasserstein variation of the conditional distributions multiplied by some constant

$$\inf_{T_S} \{R(g_B \circ T_S, X) - R(g_B, X, S)\} \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_B) \right)^{\frac{1}{2}}.$$

Although the Wasserstein barycenter was already suggested in Feldman et al. [2015], the consideration of weights is a novelty and yields in fact the good repair. We will also improve their repair procedure, which in practice did not achieve the complete fairness in terms of *statistical parity*, and we provide a generalization to higher dimensions.

Finally, we propose to set a trade-off between the quality of the classification with the repaired data and the achieving of fairness by partially changing the data with our *random repair*. It consists in introducing a proportion of contaminated data which follows the distribution of the Wasserstein barycenter. Let B be a Bernoulli variable with parameter $\lambda \in [0, 1]$, representing the amount of repair desired for X, and define for $s \in \{0, 1\}$ the randomly repaired distributions

$$\tilde{\mu}_{s,\lambda} = \mathcal{L}(BT_s(X) + (1 - B)X \mid S = s).$$

This would result in the blurring of the protected class as the level of repair increases, governed by the Bernoulli parameter. Furthermore, justifications for the random repair outperforming the existing partial method called *geometric repair* [Feldman et al., 2015], as well as a computational scheme to put it into practice are provided.

1.4 Statistical approach for fairness assessment

Many methods for imposing fairness, as well as many definitions, are based on indexes that clearly depend on the particular predictive algorithm (recall the *disparate impact* for example), when in fact very different models could be trained from the same learning sample. Furthermore, algorithms are usually inaccessible in the sense that explaining how the model is chosen may be seen too intrusive by most companies, or it may be simply not possible for many of them to change their learning procedures. To beat these shortchomings in the classification problem, we propose in chapter 4 to look for a condition on the learning sample that ensures that every classifier trained from it will be fair under the *statistical parity* criterion.

Particularly in this binary classification setting, besides the *disparate impact*, the *balanced error rate* (BER) is also a commonly used index. The link between both scores as well as the characterization of the latter in terms of the distance in total variation between the distributions $\mu_s,\ s \in \{0, 1\}$, are given in Theorem 4.2.1. Essentially, this result shows that the complete absence of bias in the training data corresponds to the total confusion between the two conditional distributions. However, certifying this equality is equivalent to the homogeneity testing problem and a goodness-of-fit test does not allow such a certification. From the statistical point of view, we can only certify that the two distributions μ_0 and μ_1 are close. Thus, in view of this result, one could be tempted to consider the testing problem

$$H_0 : d_{TV}(\mu_0, \mu_1) \geq \Delta_0 \text{ vs } H_a : d_{TV}(\mu_0, \mu_1) < \Delta_0,$$

for some small $\Delta_0 > 0$. Unfortunately, this is not feasible: there exists no uniformly consistent test for this problem, see Barron [1989]. Consequently, if we want to statistically assess that μ_0 and μ_1 are not too different, we have to choose a better metric. Hence, in this thesis we propose to use Wasserstein distances for this testing problem.

Applications of optimal transportation methods have witnessed a huge development in recent times in a variety of fields, including machine learning and image processing. The number of significant breakthroughs in the involved numerical procedures can help to understand some of the reasons for this interest. We refer to Chizat et al. [2018] for a more detailed account. In the particular field of statistical inference, despite some early contributions (see, e.g., Munk and Czado [1998], del Barrio et al. [1999a], del Barrio et al. [2005] or Freitag et al. [2007]), progress has been hindered by the lack of distributional limits [Sommerfeld and Munk, 2018].

In the second part of this thesis, we aim at contributing to the asymptotic theory of the empirical transportation cost. Precisely, in chapter 5 we provide a Central Limit Theorem for the Wasserstein distance between two empirical distributions with sizes n and m, $\mathcal{W}_p(P_n, Q_m),\ p \geq 1$, for observations on the real line (see Theorem 5.2.1)

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(P_n, Q_m) - \mathbb{E}\mathcal{W}_p^p(P_n, Q_m)) \to_w N(0, (1-\lambda)\sigma_p^2(P, Q) + \lambda\sigma_p^2(Q, P)),$$

with $\frac{n}{n+m} \to \lambda \in (0, 1)$. Note that the computation of the asymptotic variance is perfectly detailed in the corresponding chapter. In the case $p > 1$ our assumptions are sharp in terms of moments and smoothness. Also in this case, we prove results dealing with the choice of centering constants by indicating a list of sufficient conditions under which it is possible to exchange the constant $\mathbb{E}\mathcal{W}_p^p(P_n, Q_m)$ by the true value $\mathcal{W}_p^p(P, Q)$. We provide a consistent estimate of the asymptotic variance which enables to build two sample tests and confidence intervals to certify the similarity between two distributions.

In the setup of fair learning, rejecting the null with the test

$$H_0 : \mathcal{W}_p(\mu_0, \mu_1) \geq \Delta_0 \quad \text{vs} \quad H_a : \mathcal{W}_p(\mu_0, \mu_1) < \Delta_0,$$

will statistically certify that the distributions μ_0 and μ_1 are not too different. This will guarantee that the data set is fair, in the sense described above. In conclusion, we provide a new way of assessing fairness in machine learning by considering confidence intervals for the degree of dissimilarity between these distributions (with respect to the Wasserstein distance). Also, in the last section, we outline how our fairness assessment procedure can be tuned in order to use it with high-dimensional data.

Finally, we complete the asymptotic study of the empirical transportation cost proving a moderate deviation principle in general dimension in chapter 6. Exploiting the same idea of the linearization approach to obtain the CLT for the empirical quadratic transportation cost in general dimension in del Barrio and Loubes [2019], we prove some moment inequalities under more restrictive assumptions. This helps us to analyse the exponential convergence in probability of

$$\mathcal{W}_2^2(P_n, Q) - \mathbb{E}\mathcal{W}_2^2(P_n, Q)$$

towards 0, and subsequently to obtain a moderate deviation principle for such a statistic.

1.5 Deformation model for fair learning

Wasserstein barycenters and variance-like criterion using Wasserstein distance are used in many problems to analyze the homogeneity of collections of distributions and structural relationships between the observations. In chapter 7, we continue the study of the asymptotic theory of

the transportation cost with applications to the assessment of structural relationships between distributions. In particular, we propose the estimation of the quantiles of the empirical process of the Wasserstein's variation using a bootstrap procedure. Then we use these results for statistical inference on a distribution registration model for general deformation functions. The tests are based on the variance of the distributions with respect to their Wasserstein's barycenters for which we prove central limit theorems, including bootstrap versions.

The application of these results to the fair learning problem is part of the future work of this thesis. A schematic and brief idea to address it in a general setting could be the following. Consider observations $(X_1, S_1, Y_1), \dots, (X_n, S_n, Y_n)$ i.i.d. from the random vector (X, S, Y), where $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $d \geq 1$, and $S \in \mathcal{S} = \{1, \dots, k\}$ is discrete. For each $s \in \mathcal{S}$ and $i \in \{1, \dots, n\}$, let us denote by $X_{s,i} := X_i$ the observations of the usable attribute such that $S_i = s$ and by n_s the size of each protected group. We will moreover assume that the bias in the observed sample comes from the influence of the nuisance sensitive variable S, in the sense that the conditional distributions $\mu_s := \mathcal{L}(X|S = s), s \in \mathcal{S}$, are different. In this framework, we propose to explain the presence of bias in the observed sample through a deformation model for the data. That is, we will suppose that there exist some warping functions $(\varphi_0^*, \dots, \varphi_k^*)$ belonging to a family $\mathcal{G} = \mathcal{G}_0 \times \cdots \times \mathcal{G}_k$, and some random variables $\eta_{s,1}, \dots, \eta_{s,n_s}$, independent and equally distributed from a common but unknown distribution ν and such that, for every $s \in \mathcal{S}$,

$$X_{s,i} = (\varphi_s^*)^{-1}(\eta_{s,i}),\ 1 \leq i \leq n_s.$$

With this approach, the problem of reparing the data could be addressed through a deformation model since: (i) φ_S^* will be the optimal transport map pushing μ_S towards their Wasserstein barycenter μ_B, and (ii) $\tilde{X}_i := \eta_{S,i} = \varphi_S^*(X_i), i \in \{1, \dots, n\}$, will be the repaired version of the data that we are looking for.

Chapitre 1

Introduction

Matières

1.1 Motivation et cadre pour le problème de l'équité

Les technologies de l'intelligence artificielle ont sans aucun doute facilité la vie de l'homme ces dernières années. En particulier, les systèmes basés sur l'apprentissage machine atteignent la société dans son ensemble, dans de nombreux aspects de la vie quotidienne et du monde professionnel. Les voitures à conduite autonome, la reconnaissance précise du cancer sur les radiographies ou la prédiction de nos habitudes à partir de nos comportements passés ne sont que quelques exemples du large éventail d'applications technologiques dans lesquelles elles sont très prometteuses. Pourtant, en dépit de leurs avantages, les techniques d'apprentissage automatique ne sont pas complètement objectives, car les classifications et les prédictions des modèles reposent largement sur des données potentiellement biaisées. Cette généralisation des algorithmes prédictifs s'est donc accompagnée de préoccupations quant aux problèmes éthiques qui pourraient découler de l'adoption de ces technologies, non seulement au sein de la communauté des chercheurs mais aussi de la population tout entière. Ainsi, l'émergence d'approches multidisciplinaires pour évaluer et supprimer la présence de biais dans les algorithmes d'apprentissage machine a été fortement encouragée.

L'apprentissage équitable, ou *fair learning*, est un domaine d'apprentissage machine récemment créé qui étudie comment garantir que les préjugés dans les données et les inexactitudes des algorithmes ne conduisent pas à des modèles qui traitent les individus de manière défavorable sur la base de caractéristiques telles que la race, le sexe, les handicaps ou l'orientation sexuelle ou politique, pour ne citer que les plus frappantes. L'objectif de cette thèse est de présenter une approche mathématique du problème de l'équité dans l'apprentissage machine. L'application de nos résultats théoriques vise à faire la lumière sur le maelström de techniques ou de simples heuristiques qui semblent sans cesse aborder ces questions. Nous pensons qu'une base mathématique solide est cruciale pour garantir un traitement équitable à chaque sous-groupe

de population, ce qui contribuera à réduire la méfiance croissante de la société à l'égard des systèmes d'apprentissage machine.

Le cadre mathématique de l'apprentissage équitable est généralement présenté comme suit dans la littérature. Considérons l'espace de probabilité $\big(\Omega \subset \mathbb{R}^d, \mathcal{B}, \mathbb{P}\big)$, avec $\mathcal{B}$ le Borel $\sigma-$algebra des sous-ensembles de $\mathbb{R}^d,\ d \geq 1$. Nous supposerons dans ce qui suit que le biais est modélisé par la variable aléatoire $S \in \mathcal{S}$ qui représente une information sur les observations $X \in \mathcal{X} \subset \mathbb{R}^d$, qui ne doit pas être incluse dans le modèle pour la prédiction de la cible $Y \in \mathbb{R}^d,\ d \geq 1$. La variable S est appelée *l'attribut protégé* ou *sensible*, et on suppose généralement qu'elle est observée. Enfin, la classe de fonctions mesurables $f : (X, S) \mapsto Y$ sera désignée par $\mathcal{F}$ et, en particulier, $\mathcal{G}$ désignera la classe des classificateurs binaires.

1.2 Biais et définition de l'équité dans l'apprentissage machine

L'une des premières étapes consiste à comprendre comment le biais pourrait s'introduire dans les décisions automatiques. D'un point de vue mathématique, nous décrirons au chapitre 3 deux modèles possibles, proposés d'abord dans Serrurier et al. [2019], qui visent à donner un aperçu de cette question. Le premier modèle (voir Figure 3.1a) correspond au cas où les données X sont soumises à la variable de nuisance de biais S qui, en principe, est supposée ne pas être impliquée dans la tâche d'apprentissage, et dont l'influence dans la prédiction devrait être supprimée. Dans cette hypothèse, un modèle équitable exige que le résultat ne dépende pas de la variable sensible. D'autre part, le second modèle (voir Figure 3.1b) traite de la situation où une décision biaisée est observée à la suite d'un score juste qui a été biaisé par les utilisations donnant lieu à l'objectif Y. Ainsi, un modèle équitable dans ce cas modifiera la prédiction afin de la rendre indépendante de la variable protégée. Nous observons donc que la notion probabiliste sous-jacente à chaque modèle est un type différent d'indépendance entre les distributions. Le choix de cette hypothèse est donc déterminant dans le critère utilisé pour l'équité. En ce sens, nous allons examiner la notion de *perfect fairness* comme une indépendance entre la variable protégée S et le résultat $\widehat{Y} = f(X, S)$, les deux considérant de manière conditionnelle (deuxième modèle) ou non (premier modèle) la valeur réelle de la cible Y. Chaque approche a donné lieu à des définitions différentes :

- *Statistical parity* (SP) [Dwork et al., 2012] traite de $\hat{Y} \perp\!\!\!\perp S$
- *Equality of odds* (EO) [Hardt et al., 2016] considère $\hat{Y} \mid Y \perp\!\!\!\perp S$, et est particulièrement bien adapté aux scénarios où la vraie valeur est disponible pour les décisions historiques utilisées pendant la phase de formation.

La plupart des théories de l'équité ont été développées en particulier dans le cas où $S \in \mathcal{S} = \{0, 1\}$ est une variable binaire. En d'autres termes, la population est censée être éventuellement divisée en deux catégories, en prenant la valeur $S = 0$ pour la *minorité* (supposée être la classe défavorisée), et $S = 1$ pour le *default* (et généralement la classe préférée). Nous étudierons donc plus en profondeur ce cas dans la première partie de la thèse, en commençant par le chapitre 2 qui est consacré à la classification binaire. Son but est de motiver le problème de l'équité dans l'apprentissage machine en présentant quelques résultats statistiques complets obtenus à partir de l'étude du critère *statistical parity* avec les applications à la notation des crédits. Nous considérons spécifiquement l'ensemble de données réelles *Adult Income*[1]. Elle consiste à prévoir une variable binaire qui correspond à un revenu inférieur ou supérieur à 50k$ par an, où le déséquilibre existant entre la prévision de revenu et les variables binaires sensibles *Genre* et *Origine ethnique* est clairement perceptible. Cette décision pourrait éventuellement être utilisée

[1] `https://archive.ics.uci.edu/ml/datasets/adult`

pour évaluer le risque de crédit des demandeurs de prêts, ce qui rend cet ensemble de données particulièrement populaire dans la communauté de l'apprentissage automatique. Dans ce cadre, le biais d'un classificateur binaire $g(X, S) = \hat{Y}$ est fréquemment quantifié avec le *disparate impact* (DI) :

$$DI(g, X, S) = \frac{\mathbb{P}(g(X, S) = 1|S = 0)}{\mathbb{P}(g(X, S) = 1|S = 1)}.$$

Cet indice a été introduit pour la première fois sous la forme de la règle des 4/5 par la Commission des pratiques d'emploi équitables de l'État de Californie (FEPC) en 1971[2]. Depuis lors, le seuil de $0,8$ a été choisi dans différents procès comme note légale pour juger si les discriminations commises par un algorithme sont acceptables ou non (voir Feldman et al. [2015], Mercat-Bruns [2016] ou Zafar et al. [2017a]). Pourtant, ce score, ainsi que beaucoup d'autres décrits dans la littérature sur l'apprentissage équitable, sont souvent utilisés sans contrôle statistique. Dans les cas où des procédures de test ou des intervalles de confiance sont fournies, elles sont calculées en utilisant un schéma de rééchantillonnage pour obtenir des intervalles de confiance gaussiens standardisés sous une hypothèse gaussienne qui ne correspond pas à la distribution des observations. Dans ce chapitre, nous encourageons l'utilisation des intervalles de confiance pour contrôler le risque d'évaluation faussement discriminatoire. Il est important de noter que nous obtenons la distribution asymptotique exacte des estimations des différents critères d'équité en utilisant l'approche classique du Delta-méthode [Van der Vaart, 1998]. En outre, nous montrons que certaines approches standard, notamment la suppression de la variable sensible ou l'utilisation de techniques *testing*, ne sont pas pertinentes pour tenter de corriger le comportement discriminatoire des algorithmes d'apprentissage machine. Enfin, nous testerons deux solutions a priori naïves consistant soit à construire un algorithme différencié pour chaque classe de la population, soit à adapter la décision d'un algorithme unique de manière différente pour chaque sous-population. Seule cette dernière solution s'avère utile pour obtenir une classification équitable.

Pour revenir à un contexte d'apprentissage supervisé plus général, une revue des principales méthodes d'apprentissage équitable proposées dans la littérature au cours des dernières années sera présenté d'un point de vue mathématique au chapitre 3. En outre, suivant notre approche basée sur l'indépendance, nous examinerons comment construire des algorithmes équitables et les conséquences sur la dégradation de leurs performances par rapport au cas éventuellement injuste. Cela correspond au prix de l'équité. Rappelons que la performance d'un algorithme est mesurée à travers son risque défini par

$$R(f) = \mathbb{E}(\ell(Y, f(X, S))),$$

avec $\ell : (Y, \hat{Y}) \mapsto \ell(Y, \hat{Y}) \in \mathbb{R}^+$ une certaine fonction de perte. Théoriquement, un modèle équitable $f \in \mathcal{F}$ peut être obtenu en limitant la minimisation du risque à une classe équitable de modèles, à savoir, $\inf_{f \in \mathcal{F}_{\text{Fair}}} R(f)$. Cette classe $\mathcal{F}_{\text{Fair}}$ sera particulièrement dénotée par

$$\mathcal{F}_{SP} := \{f(X, S) \in \mathcal{F} \quad \text{s.t} \quad \hat{Y} \perp\!\!\!\perp S\} \qquad \text{ou} \qquad \mathcal{F}_{EO} := \{f(X, S) \in \mathcal{F} \quad \text{s.t} \quad \hat{Y}|Y \perp\!\!\!\perp S\},$$

en fonction de la notion d'équité considérée. En général, le prix de l'équité est alors calculé comme

$$\mathcal{E}_{\text{Fair}}(\mathcal{F}) := \inf_{f \in \mathcal{F}_{\text{Fair}}} R(f) - \inf_{f \in \mathcal{F}} R(f),$$

où le $\inf_{f \in \mathcal{F}} R(f)$ est connu sous le nom de risque Bayes. Ce risque excédentaire minimal sera étudié dans ce chapitre de révision, à la fois sous des hypothèses d'équité et dans deux cadres différents : la régression et la classification. D'une part, certains résultats existants sur la limite du

[2] `https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml`

prix pour l'équité comme *statistical parity* seront refondus. Nous soulignons les points principaux suivants : (i) dans le problème de la régression, nous rappelons un résultat de Le Gouic and Loubes [2020] donnant une limite inférieure pour le risque excédentaire minimal en termes de distance quadratique de Wasserstein ; (ii) dans le problème de la classification, nous anticipons le résultat avec la limite supérieure pour le risque excédentaire minimal en termes de variation de Wasserstein proposée dans l'article Gordaliza et al. [2019], que nous mentionnerons plus loin dans cette introduction puisqu'elle correspond au contenu du chapitre 4. D'autre part, le prix de l'équité comme *equality of odds* est également étudié. Il est important de noter que de nouveaux résultats donnant les expressions du classificateur de l'équité optimale et du prédicteur de l'équité optimale (sous un modèle de régression linéaire gaussien) seront présentés.

1.3 Imposer l'équité avec une méthodologie de réparation

L'importance de garantir l'équité des résultats algorithmiques a soulevé la nécessité de concevoir des procédures pour éliminer la présence potentielle de biais. D'un point de vue procédural, les méthodes permettant d'imposer l'équité peuvent être divisées en trois grandes familles. Les méthodes de la première famille consistent à pré-traiter les données ou à extraire des représentations qui ne contiennent pas de biais indésirables, qui peuvent ensuite être utilisées comme entrée dans un modèle d'apprentissage machine standard. Les méthodes de la deuxième famille, également appelées *in-processing*, visent à forcer les modèles à produire des résultats équitables en imposant des contraintes d'équité dans le mécanisme d'apprentissage. Les méthodes de la troisième famille consistent à post-traiter les résultats d'un modèle afin de les rendre équitables. Cependant, la construction de modèles équitables parfaits peut conduire à une précision médiocre : obtenir un monde équitable avec une action positive pourrait diminuer l'efficacité définie comme sa similarité avec les utilisations contrôlées par l'échantillon test. Alors que dans certains domaines d'application, il est souhaitable de garantir le plus haut niveau d'équité possible, dans d'autres, notamment les soins de santé, la justice pénale ou les applications industrielles, les performances ne devraient pas être diminuées car les décisions auraient de graves implications pour les individus et la société. Il est donc très intéressant d'établir un compromis entre l'équité et l'exactitude, ce qui entraîne un assouplissement de la notion d'équité qui est fréquemment présentée dans la littérature comme *almost* ou *approximate fairness*. À cette fin, la plupart des méthodes se rapprochent des desiderata d'équité par des exigences sur les moments d'ordre inférieur ou d'autres fonctions des distributions correspondant à différents attributs sensibles.

En particulier, dans le chapitre 4 nous présentons notre méthodologie de réparation, qui est incluse dans la première catégorie de méthodes. Là, la notion d'équité à travers le prisme de la *statistical parity* est considérée dans le cadre de la classification binaire. Notre proposition de réparation consiste à modifier la distribution initiale de la variable d'entrée X conditionnellement au groupe protégé S, désigné par $\mu_s := \mathcal{L}(X|S=s), s \in \{0,1\}$, afin de les rendre égales (*réparation totale* pour une équité parfaite) ou suffisamment proches (*réparation partielle* pour une équité quasi-totale) d'une nouvelle distribution cible inconnue. Plus précisément, *réparation totale* revient à mapper la variable originale X dans une nouvelle variable $\tilde{X} = T_S(X)$ de telle sorte que les distributions conditionnelles par rapport à S soient identiques, à savoir,

$$\mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right).$$

Notez que la transformation $T_S : \mathbb{R}^d \to \mathbb{R}^d$ est aléatoire puisqu'elle dépend de la valeur de la variable protégée S. Dans ce cas, tout classificateur g construit avec de telles informations

sera tel que $\mathcal{L}\left(g(\tilde{X}) \mid S=0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S=1\right)$, garantissant une équité totale de la règle de classification.

La distance de Wasserstein (alias Monge-Kantorovich) apparaît alors comme un outil approprié pour comparer les distributions de probabilité et se présente naturellement dans la théorie du transport optimal (nous nous référons à Villani [2009] pour une description détaillée). Pour P et Q deux mesures de probabilité sur $\mathbb{R}^d$, la distance de Wasserstein d'ordre 2 entre P et Q est définie comme

$$\mathcal{W}_2^2(P,Q) := \min_{\pi \in \Pi(P,Q)} \int \|x-y\|^2 d\pi(x,y)$$

où $\Pi(P,Q)$ l'ensemble des mesures de probabilité sur $\mathbb{R}^d \times \mathbb{R}^d$ avec les marginaux P et Q. Le fait crucial qu'il respecte la structure des données en fait un bon choix pour les procédures de réparation, car il permettra de préserver la relation entre le résultat et les données. Ainsi, ce choix suggère que la distribution de la réparation devrait être le barycentre de Wasserstein μ_B entre les distributions conditionnelles μ_s par rapport aux poids $\pi_s = \mathbb{P}(S=s)$ des classes protégées $s \in \{0,1\}$, à savoir

$$\mu_B \in argmin_{\nu \in \mathcal{P}_2} \left\{\pi_0 \mathcal{W}_2^2(\mu_0,\nu) + \pi_1 \mathcal{W}_2^2(\mu_1,\nu)\right\},$$

et que le meilleur moyen de l'atteindre est d'utiliser les plans de transport optimal $\mu_B = \mu_s \circ {T_s}^{-1}$, for $s = 0, 1$. Il ressort de la définition du barycentre de Wasserstein que cette méthode de réparation pourrait facilement être étendue au cas où S est multi-classe.

Comme mentionné précédemment, nous justifions cette approche en montrant dans le Théorème 4.3.3 une limite supérieure pour le prix de l'équité du transport vers le barycentre μ_B. Plus précisément, nous prouvons que l'excès de risque minimal lorsque l'on considère le meilleur classificateur g_B (règle de Bayes) avec les données réparées et les données originales est limité par la variation de Wasserstein pondérée des distributions conditionnelles multipliée par une constante

$$\inf_{T_S} \{R(g_B \circ T_S, X) - R(g_B, X, S)\} \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_B)\right)^{\frac{1}{2}}.$$

Bien que le barycentre de Wasserstein ait déjà été suggéré dans Feldman et al. [2015], la prise en compte des poids ainsi que le contrôle de l'erreur sont des nouveautés importantes. Nous allons également améliorer leur procédure de réparation, qui en pratique n'a pas atteint l'équité complète en termes de *statistical parity*, et nous fournissons une généralisation à des dimensions plus élevées.

Enfin, nous proposons d'établir un compromis entre la qualité de la classification avec les données réparées et la réalisation de l'équité en modifiant partiellement les données avec notre *random repair*. Cette methode consiste à introduire une proportion de données contaminées qui suit la distribution du barycentre de Wasserstein. Soit B une variable de Bernoulli avec le paramètre $\lambda \in [0,1]$, représentant la quantité de réparation souhaitée pour X, et définir pour $s \in \{0,1\}$ les distributions réparées de façon aléatoire

$$\tilde{\mu}_{s,\lambda} = \mathcal{L}(BT_s(X) + (1-B)X \mid S=s).$$

Cela conduirait à un brouillage de la classe protégée à mesure que le niveau de réparation augmente, régi par le paramètre de Bernoulli. En outre, les justifications de la réparation aléatoire surpassent la méthode partielle existante appelée *geometric repair* [Feldman et al., 2015], ainsi qu'un schéma de calcul pour le mettre en pratique sont fournis.

1.4 Approche statistique pour l'évaluation de l'équité

De nombreuses méthodes pour imposer l'équité, ainsi que de nombreuses définitions, sont basées sur des indices qui dépendent clairement de l'algorithme prédictif particulier (rappelez-vous le *disparate impact* par exemple), alors qu'en fait des modèles très différents pourraient être formés à partir du même échantillon d'apprentissage. En outre, les algorithmes sont généralement inaccessibles dans le sens où expliquer comment le modèle est choisi peut être considéré comme trop intrusif par la plupart des entreprises, ou il peut être tout simplement impossible pour beaucoup d'entre elles de modifier leurs procédures d'apprentissage. Pour surmonter ces difficultés dans le problème de la classification, nous proposons au chapitre 4 de rechercher une condition sur l'échantillon d'apprentissage qui garantisse que chaque classificateur formé à partir de celui-ci sera équitable selon le critère *statistical parity*.

En particulier, dans ce paramètre de classification binaire, outre le *disparate impact*, le *balanced error rate* (BER) est également un indice utilisé en commun. Le lien entre les deux scores ainsi que la caractérisation de ce dernier en termes de distance de variation totale entre les distributions $\mu_s,\ s \in \{0,1\}$, sont donnés dans le Théorème 4.2.1. Essentiellement, ce résultat montre que l'absence totale de biais dans les données d'aprentissage correspond à la confusion totale entre les deux distributions conditionnelles. Cependant, la certification de cette égalité équivaut au problème du test d'homogénéité et un test d'adéquation ne permet pas une telle certification. D'un point de vue statistique, on ne peut que certifier que les deux distributions μ_0 et μ_1 sont proches. Ainsi, au vu de ce résultat, on pourrait être tenté de considérer le problème du test

$$H_0 : d_{TV}\left(\mu_0, \mu_1\right) \geq \Delta_0 \text{ vs } H_a : d_{TV}\left(\mu_0, \mu_1\right) < \Delta_0,$$

pour quelque petit $\Delta_0 > 0$. Malheureusement, cela n'est pas possible : il n'existe pas de test uniformément consistent pour ce problème, voir Barron [1989]. Par conséquent, si nous voulons évaluer statistiquement que μ_0 et μ_1 ne sont pas trop différents, nous devons choisir une meilleure métrique. C'est pourquoi, dans cette thèse, nous proposons d'utiliser les distances de Wasserstein pour ce problème de test.

Les applications des méthodes de transport optimal ont connu un développement considérable ces derniers temps dans divers domaines, notamment l'apprentissage machine et le traitement de l'image. Le nombre de percées significatives dans les procédures numériques concernées peut aider à comprendre certaines des raisons de cet intérêt. Nous renvoyons à Chizat et al. [2018] pour un compte rendu plus détaillé. Dans le domaine particulier de l'inférence statistique, malgré quelques contributions précoces (voir, par exemple, Munk and Czado [1998], del Barrio et al. [1999a], del Barrio et al. [2005] ou Freitag et al. [2007]), les progrès ont été entravés par l'absence de limites de distribution [Sommerfeld and Munk, 2018].

Dans la deuxième partie de cette thèse, nous voulons contribuer à la théorie asymptotique du coût empirique du transport. Précisément, dans le chapitre 5 nous fournissons un Théorème de limite central pour la distance de Wasserstein entre deux distributions empiriques de tailles n et m, $\mathcal{W}_p(P_n, Q_m),\ p \geq 1$, pour les observations sur la droite réelle (voir le Théorème 5.2.1)

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(P_n, Q_m) - \mathbb{E}\mathcal{W}_p^p(P_n, Q_m)) \to_w N(0, (1-\lambda)\sigma_p^2(P, Q) + \lambda\sigma_p^2(Q, P)),$$

où $\frac{n}{n+m} \to \lambda \in (0,1)$. Notez que le calcul de la variance asymptotique est parfaitement détaillé dans le chapitre correspondant. Dans le cas $p > 1$, nos hypothèses sont minimales en termes de moments et de régularité. Dans ce cas également, nous traitons du choix des constantes de centrage, en indiquant une liste de conditions suffisantes dans lesquelles il est possible d'échanger la constante $\mathbb{E}\mathcal{W}_p^p(P_n, Q_m)$ par la vraie valeur $\mathcal{W}_p^p(P, Q)$. Nous fournissons une estimation

consistent de la variance asymptotique qui permet de construire deux tests d'échantillons et des intervalles de confiance pour certifier la similarité entre deux distributions.

Dans la mise en place d'un apprentissage équitable, rejeter le nul avec le test

$$H_0 : \mathcal{W}_p(\mu_0, \mu_1) \geq \Delta_0 \quad \text{vs} \quad H_a : \mathcal{W}_p(\mu_0, \mu_1) < \Delta_0,$$

certifiera statistiquement que les distributions μ_0 et μ_1 ne sont pas trop différentes. Cela garantira que l'ensemble de données est équitable, au sens décrit ci-dessus. En conclusion, nous proposons une nouvelle façon d'évaluer l'équité de l'apprentissage machine en considérant les intervalles de confiance pour le degré de dissimilitude entre ces distributions (par rapport à la distance de Wasserstein). Dans la dernière section, nous expliquons comment notre procédure d'évaluation de l'équité peut être ajustée pour être utilisée avec des données de grande dimension.

Enfin, nous complétons l'étude asymptotique du coût de transport empirique prouvant un principe des déviations modérées en dimension générale au chapitre 6. En explitant la même idée de l'approche de linéarisation pour obtenir le TLC pour le coût empirique quadratique de transport en del Barrio and Loubes [2019], nous prouvons des inégalités de moment sous des hypothèses plus restrictives. Cela nous aide à analyser la convergence exponentielle en probabilité de

$$\mathcal{W}_2^2(P_n, Q) - \mathbb{E}\mathcal{W}_2^2(P_n, Q)$$

vers 0, et à obtenir ensuite un principe des déviations modérées pour cette statistique.

1.5 Modèle de déformation pour un apprentissage équitable

Les barycentres de Wasserstein et les critères de variance utilisant la distance de Wasserstein sont utilisés dans de nombreux problèmes pour analyser l'homogénéité des collections de distributions et les relations structurelles entre les observations. Dans le chapitre 7, nous poursuivons l'étude de la théorie asymptotique du coût de transport avec des applications à l'évaluation des relations structurelles entre les distributions. En particulier, nous proposons l'estimation des quantiles du processus empirique de la variation de Wasserstein en utilisant une procédure bootstrap. Ensuite, nous utilisons ces résultats pour l'inférence statistique sur un modèle d'enregistrement de la distribution pour les fonctions de déformation générale. Les tests sont basés sur la variance des distributions par rapport à leurs barycentres de Wasserstein pour lesquels nous prouvons les théorèmes de limite centrale, y compris les versions bootstrap.

L'application de ces résultats au problème de l'apprentissage équitable fait partie du futur travail de cette thèse. Voici un schéma et une brève idée pour l'aborder dans un cadre général. Considérer les observations $(X_1, S_1, Y_1), \ldots, (X_n, S_n, Y_n)$ i.i.d. du vecteur aléatoire (X, S, Y), où $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $d \geq 1$, et $S \in \mathcal{S} = \{1, \ldots, k\}$ est discret. Pour chaque $s \in \mathcal{S}$ et $i \in \{1, \ldots, n\}$, dénotons par $X_{s,i} := X_i$ les observations de l'attribut utilisable tel que $S_i = s$ et par n_s la taille de chaque groupe protégé. Nous supposerons en outre que le biais dans l'échantillon observé provient de l'influence de la variable sensible aux nuisances S, en ce sens que les distributions conditionnelles $\mu_s := \mathcal{L}(X|S = s), s \in \mathcal{S}$, sont différentes. Dans ce cadre, nous proposons d'expliquer la présence de biais dans l'échantillon observé par un modèle de déformation des données. C'est-à-dire que nous supposerons qu'il existe certaines fonctions de déformation $(\varphi_0^*, \ldots, \varphi_k^*)$ appartenant à une famille $\mathcal{G} = \mathcal{G}_0 \times \cdots \times \mathcal{G}_k$, et certaines variables aléatoires $\eta_{s,1}, \ldots, \eta_{s,n_s}$, indépendantes et également réparties à partir d'une distribution commune mais inconnue ν et telles que, pour chaque $s \in \mathcal{S}$,

$$X_{s,i} = (\varphi_s^*)^{-1}(\eta_{s,i}), \ 1 \leq i \leq n_s.$$

Avec cette approche, nous pouvons traiter le problème de la réparation des données comme un modèle de déformation, car on aura cela : (i) φ_S^* est la carte de transport optimale poussant μ_S vers leur barycentre de Wasserstein μ_B, et (ii) $\tilde{X}_i := \eta_{S,i} = \varphi_S^*(X_i), i \in \{1, \ldots, n\}$, sont la version réparée des données que nous recherchons.

Capítulo 1

Introducción

Contenidos

1.1 Motivación y marco del problema de la equidad

En los últimos años, las tecnologías basadas en la Inteligencia Artificial están haciendo indudablemente la vida humana más fácil. En particular, los sistemas basados en el aprendizaje automático, o en inglés *machine learning*, están alcanzando a la sociedad en general, tanto en el mundo profesional como en muchos aspectos de la vida cotidiana. Potenciar los coches autodirigidos, reconocer con precisión el cáncer en las radiografías, o predecir nuestros intereses basados en comportamientos pasados, son sólo algunos ejemplos en la amplia gama de aplicaciones en las que estas tecnologías están mostrando ser de gran valor y utilidad. Sin embargo, con todos sus beneficios, las técnicas de machine learning no son absolutamente objetivas, pues las clasificaciones y predicciones hechas por los modelos dependen en gran medida de datos potencialmente sesgados. En consecuencia, este uso generalizado de los algoritmos de predicción ha ido acompañado de preocupaciones sobre las cuestiones éticas que pueden surgir de la adopción de estas tecnologías, no sólo entre la comunidad investigadora sino también entre toda la población. Gracias a ello, se ha dado un gran impulso a la aparición de enfoques multidisciplinares para detectar y eliminar la presencia de sesgos en las decisiones automáticas tomadas por algoritmos.

El aprendizaje justo, del inglés *fair learning*, es un área recientemente establecida del aprendizaje automático que estudia cómo asegurar que los sesgos en los datos y las inexactitudes de los algoritmos no conduzcan a modelos que traten desfavorablemente a los individuos en base a características como la raza, el género, las discapacidades o la orientación sexual o política, sólo por nombrar las de mayor impacto en la opinión pública. El propósito de esta tesis es presentar un enfoque matemático del problema de la equidad en el aprendizaje automático. La aplicación de nuestros resultados teóricos tiene como objetivo arrojar un poco de luz sobre la vorágine de técnicas o meras heurísticas que aparecen incesantemente para tratar de dar respuesta a estos problemas. Creemos que una base matemática robusta es crucial para garantizar un tratamiento

justo para cada subgrupo de población, lo que contribuirá a reducir la creciente desconfianza de la sociedad hacia los sistemas de aprendizaje automático.

El marco matemático para el problema del aprendizaje justo se suele presentar en la literatura como sigue. Consideremos el espacio de probabilidad $\left(\Omega \subset \mathbb{R}^d, \mathcal{B}, \mathbb{P}\right)$, con $\mathcal{B}$ la $\sigma-$algebra de subconjuntos de $\mathbb{R}^d, d \geq 1$. Asumiremos que la variable aleatoria $S \in \mathcal{S}$ modela el sesgo, de manera que representa una información sobre las observaciones $X \in \mathcal{X} \subset \mathbb{R}^d$ que no debe ser incluida en el modelo para la predicción de la respuesta $Y \in \mathbb{R}^d,\ d \geq 1$. Esta variable S, cuyo valor se asume conocido, recibe el nombre de *atributo protegido* o *sensible* (*protected* o *sensitive attribute*, en inglés). Finalmente, denotaremos por $\mathcal{F}$ la clase de funciones medibles $f : (X, S) \mapsto Y$ y, más concretamente, por $\mathcal{G}$ cuando se trate de clasificadores binarios.

1.2 Sesgo y definición de equidad en el aprendizaje automático

Uno de los primeros pasos es motivar la importancia de comprender cómo los sesgos se pueden introducir en las decisiones automáticas. Desde un punto de vista matemático, describiremos en el capítulo 3 dos posibles modelos, propuestos por primera vez en Serrurier et al. [2019], que tienen como objetivo formalizar esta cuestión. El primer modelo (véase la Figura 3.1a) corresponde al caso en que los datos X están sujetos a la variable del sesgo S que, en principio, se supone que no está involucrada en la tarea de aprendizaje, y cuya influencia en la predicción debe ser eliminada. Bajo este supuesto, un modelo justo requiere que el resultado no dependa de esta variable sensible. Por otra parte, el segundo modelo (véase la Figura 3.1b) se ocupa de la situación en que se observa una decisión sesgada como resultado de una puntuación justa que ha sido sesgada por los usos que dan lugar al objetivo Y. En este caso, un modelo justo cambiaría la predicción para hacerla independiente de la variable protegida. Observamos que la noción probabilística que subyace en cada modelo es un tipo diferente de independencia entre distribuciones. Por lo tanto, la elección de esta hipótesis es decisiva en el criterio utilizado para garantizar la equidad. En este sentido, consideraremos la noción de *equidad perfecta* (*perfect fairness*) como una independencia entre la variable protegida S y el resultado $\widehat{Y} = f(X, S)$, tanto condicionalmente al verdadero valor del objetivo Y (segundo modelo) como no condicionalmente (primer modelo). Cada uno de estos enfoques ha dado lugar a diferentes definiciones:

- *Statistical parity* (SP) [Dwork et al., 2012] trata con $\hat{Y} \perp\!\!\!\perp S$
- *Equality of odds* (EO) [Hardt et al., 2016] considera $\hat{Y}|Y \perp\!\!\!\perp S$, y es especialmente adecuada para los escenarios en los que se dispone del verdadero valor de la etiqueta objetivo para las decisiones históricas utilizadas durante la fase de entrenamiento.

La mayor parte de la teoría de la equidad se ha desarrollado particularmente en el caso de que $S \in \mathcal{S} = \{0, 1\}$ es una variable binaria. Es decir, se supone que la población se encuentra dividida en dos categorías, tomando el valor $S = 0$ para la minoría (supuestamente la clase desfavorecida) y $S = 1$ para la mayoría o clase por defecto (supuestamente la clase favorecida). Por tanto, estudiaremos este caso de manera más detallada en la primera parte de la tesis, empezando en el capítulo 2, el cual está enmarcado en este escenario concreto de clasificación binaria. Su objetivo es motivar el problema de la equidad en el aprendizaje automático mediante la presentación de resultados completos del estudio del criterio *statistical parity* con aplicaciones en puntuación crediticia. Específicamente, consideramos la base de datos reales conocido como *Adult Income*[1]. Este conjunto consiste en predicciones de la variable binaria correspondiente a si un individuo tiene ingresos anuales superiores a 50.000 $, en las que claramente se aprecia un desequilibrio

[1] `https://archive.ics.uci.edu/ml/datasets/adult`

entre aquellas hechas para individuos con distinto género y origen étnico. El hecho de que esta predicción pueda ser potencialmente utilizada para evaluar el riesgo de los solicitantes de créditos, ha popularizado este conjunto de datos entre la comunidad del aprendizaje automático. En este contexto de clasificación binaria, es frecuente cuantificar el sesgo de un clasificador $g(X, S) = \hat{Y}$ mediante el *disparate impact* (DI):

$$DI(g, X, S) = \frac{\mathbb{P}(g(X, S) = 1|S = 0)}{\mathbb{P}(g(X, S) = 1|S = 1)}.$$

Este índice se introdujo como la regla de los 4/5 en el *State of California Fair Employment Practice Commission (FEPC)* en 1971[2]. Desde entonces, en numerosos juicios se ha elegido el umbral 0,8 para aprobar la equidad en las decisiones algorítmicas (véase por ejemplo Feldman et al. [2015], Mercat-Bruns [2016] o Zafar et al. [2017a]). Sin embargo, este score, al igual que la mayoría de los descritos en la literatura, se utiliza frecuentemente sin un control estadístico. Además, en muchos casos, los contrastes de hipótesis o regiones de confianza son obtenidos bajo hipótesis de normalidad que no se corresponde con la distribución de las observaciones. En este capítulo 2, proponemos el uso de intervalos de confianza para controlar el riesgo de falsedad en las evaluaciones discriminatorias. Cabe destacar la obtención en esta tesis de la distribución asintótica de varios criterios de equidad, a través del clásico Delta-método [Van der Vaart, 1998]. Además, mostramos cómo algunos de los procedimientos estándar, tales como la eliminación del valor de la variable sensible de la muestra de entrenamiento o las técnicas llamadas *testing* (detalladas más adelante), no son en absoluto efectivos cuando se trata de corregir los comportamientos discriminatorios de los algoritmos. Finalmente, comprobamos dos soluciones que consisten o bien en construir algoritmos diferenciados para cada clase, o bien en adaptar la decisión de un mismo algoritmo a cada clase; y concluimos que sólo la segunda obtiene clasificaciones justas.

Volviendo a un contexto de aprendizaje supervisado más general, en el capítulo 3 se hace una revisión de las principales metodologías de aprendizaje justo que se han propuesto en lo últimos años. Además, plantearemos cómo construir algoritmos justos y cómo valorar la consecuente degradación en su desempeño, en comparación con el caso posiblemente injusto. Esta cuestión corresponde con lo que se suele denominar *precio de la equidad*. Recordemos que la eficiencia de un algoritmo se mide mediante el riesgo definido por

$$R(f) = \mathbb{E}(\ell(Y, f(X, S))),$$

con $\ell : (Y, \hat{Y}) \mapsto \ell(Y, \hat{Y}) \in \mathbb{R}^+$ cierta función de pérdida. Teóricamente, un modelo justo $f \in \mathcal{F}$ se obtiene como resultado de la minimización del riesgo dentro de una clase de modelos justos, es decir, $\text{ínf}_{f \in \mathcal{F}_{\text{Fair}}} R(f)$. En particular, denotaremos esta clase $\mathcal{F}_{\text{Fair}}$ por

$$\mathcal{F}_{SP} := \{f(X, S) \in \mathcal{F} \quad \text{s.t} \quad \hat{Y} \perp\!\!\!\perp S\} \qquad \text{o} \qquad \mathcal{F}_{EO} := \{f(X, S) \in \mathcal{F} \quad \text{s.t} \quad \hat{Y}|Y \perp\!\!\!\perp S\},$$

dependiendo de la noción considerada de equidad. En general, el precio de la equidad se calcula entonces como

$$\mathcal{E}_{\text{Fair}}(\mathcal{F}) := \inf_{f \in \mathcal{F}_{\text{Fair}}} R(f) - \inf_{f \in \mathcal{F}} R(f),$$

donde $\text{ínf}_{f \in \mathcal{F}} R(f)$ se conoce como el riesgo de Bayes. En este capítulo de revisión se estudiará este exceso de riesgo mínimo, bajo ambas nociones de equidad y en dos marcos diferentes: regresión y clasificación. Por un lado, revisitaremos algunos resultados existentes sobre la acotación

[2] `https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml`

del precio de la equidad como *statistical parity*. En particular, (i) en el problema de regresión, destacamos el resultado de Le Gouic and Loubes [2020] que proporciona una cota inferior para el exceso de riesgo mínimo en términos de la distancia cuadrática de Wasserstein; (ii) en el problema de clasificación, anticipamos la cota superior para el exceso de riesgo mínimo en términos de la variación de Wasserstein, propuesta en el trabajo Gordaliza et al. [2019], a la cual haremos referencia más adelante en esta introducción, puesto que corresponde al contenido del capítulo 4. Por otro lado, también estudiamos el precio de la equidad como *equality of odds*. En este caso, obtenemos las expresiones exactas del clasificador y predictor (bajo un modelo de regresión normal) óptimos y justos.

1.3 Una nueva metodología de reparación para imponer equidad

La importancia de asegurar la equidad en los resultados algorítmicos ha suscitado la necesidad de diseñar procedimientos para eliminar la presencia potencial de sesgos. Desde el punto de vista del procedimiento, los métodos para imponer la equidad se dividen habitualmente a grandes rasgos en tres familias. En primer lugar, existe una familia de métodos que consisten en un preprocesado de los datos o en la extracción de representantes libres de sesgos indeseados, los cuales pueden ser posteriormente utilizados como input en un modelo de machine learning estándar. En la segunda familia, se incluyen los métodos que fuerzan al modelo a producir resultados justos mediante la imposición de restricciones al mecanismo de aprendizaje. Por último, los métodos en la tercera familia consisten en un post-procesado del resultado de la predicción del modelo con el objetivo de hacerlo justo. Sin embargo, construir modelos perfectamente equitativos puede conducir a una pérdida notable en su exactitud: tratar de cambiar el mundo con buenas intenciones puede dañar la eficiencia de los modelos, entendida como la similaridad a los usos monitorizados a través de la muestra de entrenamiento. Mientras que en algunos campos de aplicación es deseable alcanzar el nivel más alto de equidad posible, en otros, tales como la sanidad o la justicia penal, la eficiencia no debe ser disminuida, pues las decisiones pueden tener implicaciones muy graves sobre la vida de las personas o la sociedad en general. Por lo tanto, es de gran interés establecer un equilibrio entre equidad y eficiencia de los modelos. Esto ha llevado a una relajación de la noción de equidad que se presenta frecuentemente en la literatura como equidad aproximada (en inglés *almost* o *approximate fairness*). Con este propósito, la mayoría de los métodos aproximan la noción de equidad mediante requerimientos sobre los momentos de orden bajo o sobre otras funciones de las distribuciones de los datos X o de la predicción $\widehat{Y}$ condicionadas a los atributos protegidos.

En particular, en el capítulo 4 presentamos nuestra metodología de reparación, que se incluye en la primera categoría de procedimientos para imponer equidad. Para ello, consideramos el criterio de *statistical parity* en el marco de la clasificación binaria. Nuestra propuesta de reparación consiste en modificar las distribuciones originales de la variable aleatoria de entrada X condicionadas al valor del atributo protegido S, denotadas como $\mu_s := \mathcal{L}\left(X|S=s\right), s \in \{0,1\}$, con el objetivo de hacerlas idénticas (*reparación total* para una equidad perfecta) o acercarlas lo suficiente (*reparación parcial* para un equidad aproximada) a una distribución nueva y desconocida. Formalmente, hacer una *reparación total* significa transformar la variable original X en la nueva $\tilde{X} = T_S(X)$, de tal manera que las distribuciones condicionadas con respecto a S coincidan

$$\mathcal{L}\left(\tilde{X} \mid S=0\right) = \mathcal{L}\left(\tilde{X} \mid S=1\right).$$

Notemos el carácter aleatorio de la transformación $T_S : \mathbb{R}^d \to \mathbb{R}^d$, pues depende del valor de S. Como resultado, cualquier clasificador g construido a partir de la nueva información satisfará $\mathcal{L}\left(g(\tilde{X}) \mid S=0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S=1\right)$, garantizando la equidad completa.

La distancia de Wasserstein (también conocida como de Monge-Kantorovich) surge de manera natural del problema del transporte óptimo y ha demostrado ser una herramienta apropiada para comparar distribuciones de probabilidad (nos referimos a Villani [2009] para una descripción detallada). La distancia cuadrática de Wasserstein entre dos medidas P y Q se define como

$$\mathcal{W}_2^2(P,Q) := \min_{\pi\in\Pi(P,Q)} \int \|x-y\|^2 d\pi(x,y),$$

donde $\Pi(P,Q)$ denota el conjunto de medidas en el espacio producto $\mathbb{R}^d \times \mathbb{R}^d$ con marginales P y Q. Uno de sus rasgos característicos es que respeta la estructura de los datos, lo que la hace especialmente adecuada para los procedimientos de reparación, pues conservará la relación existente entre la respuesta y los datos de entrada originales. Por tanto, esta elección sugiere por un lado, que la distribución de la reparación sea el baricentro de Wasserstein μ_B entre las distribuciones condicionales μ_s con respecto a los pesos de las clases protegidas $\pi_s = \mathbb{P}(S = s),\ s \in \{0,1\}$, formalmente

$$\mu_B \in argmin_{\nu\in\mathcal{P}_2} \left\{\pi_0 \mathcal{W}_2^2(\mu_0,\nu) + \pi_1 \mathcal{W}_2^2(\mu_1,\nu)\right\},$$

y por otro lado, que la manera óptima de alcanzarlo sean los planes de transporte óptimo $\mu_B = \mu_s \circ {T_s}^{-1},$ for $s = 0, 1$. Además, notemos que de la definición de baricentro se deduce que la metodología de reparación propuesta es fácilmente extensible al caso en que el atributo protegido es discreto con más de dos clases.

Como ya se ha mencionado anteriormente, este enfoque está justificado en el Teorema 4.3.3 donde se proporciona una cota superior para el precio de la equidad que se consigue mediante el transporte hacia el baricentro μ_B. Más concretamente, probamos que el exceso mínimo al comparar el riesgo del mejor clasificador g_B (regla de Bayes) con los datos reparados y con los originales está controlado por la variación ponderada de Wasserstein de las distribuciones condicionadas multiplicada por cierta constante

$$\inf_{T_S} \{R(g_B \circ T_S, X) - R(g_B, X, S)\} \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s,\mu_B)\right)^{\frac{1}{2}}.$$

A pesar de que el baricentro de Wasserstein ya se había sugerido con anterioridad en trabajos como el de Feldman et al. [2015], la consideración de los pesos es una novedad de nuestro resultado y conduce a la buena reparación de los datos. Además, mejoramos el esquema computacional propuesto en el mencionado trabajo, el cual en la práctica no alcanza la equidad completa en términos de *statistical parity*, y proporcionamos una manera de generalizarlo a altas dimensiones.

Finalmente, proponemos una metodología de reparación parcial a la que denominamos *random repair*, que pretende establecer un equilibrio entre el nivel alcanzado de equidad y la calidad de la clasificación que resulta de los datos reparados. Este método consiste en introducir una proporción de datos contaminados que siguen la ley del baricentro μ_B. Para ello, denotemos por B una variable Bernoulli con parámetro $\lambda \in [0,1]$, que representa la cantidad de reparación deseada para X. Definimos para cada $s \in \{0,1\}$ las distribuciones aleatoriamente reparadas como

$$\tilde{\mu}_{s,\lambda} = \mathcal{L}(BT_s(X) + (1-B)X \mid S = s).$$

Como resultado, se consigue difuminar el valor de la variable protegida a medida que el nivel de reparación aumenta, gobernado por el parámetro de Bernoulli. Por último, justificamos por qué este método es mejor que uno de los esquemas de reparación parcial más conocidos en la literatura, denominado *geometric repair* [Feldman et al., 2015], y proponemos un esquema computacional para llevarlo a la práctica.

1.4 Un enfoque estadístico para la evaluación de la equidad algorítmica

Muchos de los métodos para garantizar la equidad, así como muchas de sus definiciones, están basados en índices que dependen claramente del algoritmo predictivo en cuestión (recuérdese por ejemplo el *disparate impact*) cuando, en realidad, a partir de la misma muestra se pueden entrenar modelos muy diferentes. Por otro lado, es habitual que los algoritmos sean inaccesibles, en el sentido de que las empresas podrían interpretar como intrusivo el hecho de explicar cómo construyen sus modelos, o simplemente no están interesadas en cambiarlos. Para hacer frente a estos problemas en el ámbito de la clasificación binaria, en el capítulo 4 proponemos buscar una condición sobre la muestra de aprendizaje que asegure que cualquier clasificador entrenado a partir de ella será justo en el sentido dado por el criterio de *statistical parity*.

Particularmente en este contexto, además del *disparate impact*, otro índice habitual es el llamado *balanced error rate* (BER). Dos de las contribuciones en este capítulo consisten tanto en establecer el enlace entre ambos índices, como en caracterizar el segundo en términos de la distancia en variación total entre las distribuciones $\mu_s,\ s \in \{0,1\}$. Esencialemente, en el Teorema 4.2.1 mostramos que la ausencia absoluta de sesgo en el conjunto de entrenamiento corresponde con la confusión total entre dichas distribuciones condicionadas. Sin embargo, comprobar tal igualdad equivale a un problema de homogeneidad entre distribuciones, y un test de bondad de ajuste no permite tal certificación. Desde un punto de vista estadístico, sólamente se puede certificar que las dos distribuciones μ_0 y μ_1 están cerca. Como consecuencia de este resultado, estaríamos tentados a considerar el contraste de hipótesis

$$H_0 : d_{TV}\left(\mu_0, \mu_1\right) \geq \Delta_0 \text{ vs } H_a : d_{TV}\left(\mu_0, \mu_1\right) < \Delta_0,$$

para cierto $\Delta_0 > 0$ pequeño. Desafortunadamente, esto no es viable al no existir tests uniformemente consistentes para este problema (véase en Barron [1989]). Así pues, para comprobar estadísticamente la diferencia entre μ_0 y μ_1 debemos considerar otra métrica y, en esta tesis, proponemos emplear las distancias de Wasserstein.

Recientemente, las aplicaciones de los métodos de transporte óptimo han experimentado un enorme avance en una gran cantidad de campos, tales como el machine learning o el procesado de imágenes, por citar dos de los más candentes. El creciente interés por estos métodos viene de las mejoras en los procedimientos numéricos involucrados. Para más detalles sobre este aspecto, nos referimos a Chizat et al. [2018]. Particularmente en el campo de la inferencia estadística, a pesar de algunas contribuciones tempranas en Munk and Czado [1998], del Barrio et al. [1999a], del Barrio et al. [2005] or Freitag et al. [2007], este progreso se ha visto frenado por la falta de resultados sobre distribuciones límite [Sommerfeld and Munk, 2018].

En la segunda parte de la tesis, nuestro objetivo es contribuir a la teoría asintótica del coste empírico de transporte. En concreto, en el capítulo 5 proporcionamos un teorema central del límite para la distancia de Wasserstein $\mathcal{W}_p(P_n, Q_m)$, con coste de orden $p \geq 1$, entre dos distribuciones empíricas de distintos tamaños n y m, a partir de observaciones en la recta real (véase el Teorema 5.2.1)

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(F_n, G_m) - E\mathcal{W}_p^p(F_n, G_m)) \to_w N(0, (1-\lambda)\sigma_p^2(F,G) + \lambda\sigma_p^2(G,F)),$$

siendo $\frac{n}{n+m} \to \lambda \in (0,1)$. El cálculo de la varianza asintótica se detalla perfectamente en el correspondiente capítulo. En el caso $p > 1$, las hipótesis requeridas son minimales en términos de momentos y suavidad de las distribuciones. También en este caso tratamos con la elección de las constantes de centramiento, indicando un conjunto de condiciones suficientes bajo las cuales es posible intercambiar $E\mathcal{W}_p^p(F_n, G_m))$ por el verdadero valor $\mathcal{W}_p^p(F,G)$. Por último,

proporcionamos un estimador consistente de la varianza asintótica que, bajo las mencionadas condiciones, nos permite construir tanto un test de dos muestras, como intervalos de confianza para certificar la similaridad entre dos distribuciones.

En el contexto del aprendizaje justo, podemos decir de manera coloquial que rechazar la hipótesis nula del contraste

$$H_0 : \mathcal{W}_p(\mu_0, \mu_1) \geq \Delta_0 \quad \text{vs} \quad H_a : \mathcal{W}_p(\mu_0, \mu_1) < \Delta_0,$$

certificará estadísticamente que las distribuciones μ_0 y μ_1 no son demasiado diferentes. Esto garantizará la equidad en el conjunto de datos, en el sentido indicado previamente. En conclusión, proporcionamos una nueva metodología de evaluación de la equidad en el aprendizaje automático basada en intervalos de confianza para el grado de disimilaridad entre estas distribuciones (con respecto a la distancia de Wasserstein). En la última sección, indicamos cómo este método puede modificarse para aplicarlo con datos en altas dimensiones.

Finalmente, el capítulo 6 completa el estudio de la teoría asintótica del coste empírico de transporte con un principio de desviaciones moderadas en dimensión general. Explotando la misma idea de la linealización para obtener el TCL para el coste empírico cuadrático de transporte en del Barrio and Loubes [2019], probamos algunos resultados sobre desigualdades de momentos, bajo ciertas condiciones más restrictivas. Tales resultados nos ayudan a analizar la convergencia exponencial en probabilidad de

$$\mathcal{W}_2^2(P_n, Q) - \mathbb{E}\mathcal{W}_2^2(P_n, Q)$$

hacia 0, y posteriormente a obtener un principio de desviaciones moderadas para este estadístico.

1.5 Modelo de deformación para el aprendizaje justo

En muchos problemas que requieren analizar la homogeneidad de una colección de distribuciones y las relaciones estructurales entre las observaciones, es habitual el empleo de los baricentros de Wassertein y de criterios de varianza basados en la distancia de Wasserstein. En el capítulo 7, continuamos con el estudio de la teoría asintótica del coste de transporte con aplicaciones a la evaluación de las relaciones estructurales existentes entre distribuciones. En particular, proponemos un procedimiento tipo *bootstrap* para estimar los cuantiles del proceso empírico de la variación de Wasserstein. Estos resultados son empleados para hacer inferencia estadística en un modelo general de deformación para distribuciones. Los tests se basan en la varianza de las distribuciones con respecto a su baricentro de Wasserstein, para la cual probamos teoremas centrales del límite, con versiones *bootstrap* incluidas.

La aplicación de estos resultados al problema de aprendizaje justo es parte del trabajo futuro de esta tesis. De manera breve, un esquema para abordar esta cuestión podría ser el siguiente. Consideremos observaciones $(X_1, S_1, Y_1), \ldots, (X_n, S_n, Y_n)$ i.i.d. del vector aleatorio (X, S, Y), donde $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $d \geq 1$, y $S \in \mathcal{S} = \{1, \ldots, k\}$ es discreta. Para cada $s \in \mathcal{S}$ y $i \in \{1, \ldots, n\}$, denotemos por $X_{s,i} := X_i$ las observaciones del atributo legítimo y utilizable tales que $S_i = s$, y por n_s el tamaño de cada uno de los grupos protegidos. Asumiremos además que el sesgo en la muestra observada procede de la influencia de la información sensible dada en S, en el sentido de que las distribuciones $\mu_s := \mathcal{L}(X|S = s), s \in \mathcal{S}$, son distintas. En este contexto, proponemos explicar dicha presencia de sesgo a través de un modelo de deformación para los datos. Esto es, supondremos que existen funciones de deformación $(\varphi_0^*, \ldots, \varphi_k^*)$ que pertenecen a una familia, en principio general, $\mathcal{G} = \mathcal{G}_0 \times \cdots \times \mathcal{G}_k$, y ciertas variables aleatorias $\eta_{s,1}, \ldots, \eta_{s,n_s}$, independientes e igualmente distribuidas que una medida desconocida ν y tales que, para cada $s \in \mathcal{S}$,

$$X_{s,i} = (\varphi_s^*)^{-1}(\eta_{s,i}), \ 1 \leq i \leq n_s.$$

Con este enfoque, podemos tratar el problema de la reparación de los datos como un modelo de deformación, ya que tendremos que: (i) φ_S^* serán los planes de transporte óptimo que llevan μ_S hacia el baricentro de Wasserstein μ_B, y (ii) $\tilde{X}_i := \eta_{S,i} = \varphi_S^*(X_i), i \in \{1, \ldots, n\}$, serán las versiones reparadas de los datos que estamos buscando.

Part I

Fairness in Machine Learning

Chapter 2

A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set

The content of this chapter is available online in Besse et al. [2020] and currently submitted for publication. We have also provided a companion notebook at `https://github.com/XAI-ANITI/StoryOfBias/blob/master/StoryOfBias.ipynb`.

Contents

Applications based on Machine Learning models have now become an indispensable part of the everyday life and the professional world. A critical question then recently arised among the population: Do algorithmic decisions convey any type of discrimination against specific groups

of population or minorities? In this paper, we show the importance of understanding how a bias can be introduced into automatic decisions. We first present a mathematical framework for the fair learning problem, specifically in the binary classification setting. We then propose to quantify the presence of bias by using the standard *Disparate Impact* index on the real and well-known *Adult income* data set. Finally, we check the performance of different approaches aiming to reduce the bias in binary classification outcomes. Importantly, we show that some intuitive methods are ineffective. This sheds light on the fact that trying to make fair machine learning models may be a particularly challenging task, in particular when the training observations contain a bias.

2.1 Introduction

Fairness has become one of the most popular topics in machine learning over the last years and the research community is investing a large amount of effort in this area. The main motivation is the increasing impact that the lives of Human beings are experiencing due to the generalization of machine learning systems in a wide variety of fields. Originally designed to improve recommendation systems in the internet industry, they are now becoming an inseparable part of our daily lives since more and more companies start integrating Artificial Intelligence (AI) into their existing practice or products. While some of these quotidian uses may involve leisure, with vain consequences (Amazon or Netflix use recommender systems to present a customized page that offers their products according to the order of preference of each user), other ones entail particularly sensitive decisions such as in Medicine, where patient suitability for treatment is considered; in Human Resources, where candidates are sorted out on an algorithmic decision basis; in the Automotive industry, with the release of self-driving cars; in the Banking and Insurance industry, which characterize customers according to a risk index; in Criminal justice, where the COMPAS algorithm is used in the United States for recidivism prediction... For a more detailed background on these facts see for instance Romei and Ruggieri [2014b], Berk et al. [2018] Pedreschi et al. [2012] or Friedler et al. [2019], and references therein.

The technologies that AI offers certainly make life easier. It is however a common misconception that they are absolutely objective. In particular, machine learning algorithms which are meant to automatically take accurate and efficient decisions that mimic and even sometimes outmatch human expertise, rely heavily on potentially biased data. It is interesting to remark that this bias is often due to an inherent social bias existing in the population that is used to generate the training dataset of the machine learning models. A list of potential causes for the discriminatory behaviours that machine learning algorithms may exhibit, in the sense that groups of population are treated differently, is given in Barocas and Selbst [2016]. Various real and striking cases that can be found in the literature are the following. In Angwin et al. [2016], it was found that the algorithm COMPAS used for recidivism prediction produces much higher rate of false positive predictions for black people than for white people. Later in Lahoti et al. [2019], a job platform similar to Linkedin called XING was found to predict less highly ranked qualified male candidates than female candidates. Publicly available commercial face recognition online services provided by Microsoft, Face++, and IBM respectively were also recently found to suffer from achieving much lower accuracy on females with darker skin color in Buolamwini and Gebru [2018]. Although a discrimination may appear naturally and could be thought as acceptable, as in Kamiran et al. [2010] for instance, quantifying the effect of a machine learning predictor with respect to a given situation is of high importance. Therefore, the notion of fairness in machine learning algorithms has received a growing interest over the last years. We believe this is crucial in order to guarantee a fair treatment for every subgroup of population, which will contribute to reduce the growing distrust of machine learning systems in the society.

Yet providing a definition of fairness or equity in machine learning is a complicated task and several propositions have been formulated. First described in terms of law [Winrow and Schieber, 2009], fairness is now quantified in order to detect biased decisions from automatic algorithms. We will focus on the issue of biased training data, which is one of the several possible causes of such discriminatory outcomes in machine learning mentioned above. In the fair learning literature, fairness is often defined with respect to selected variables, which are commonly denoted *protected* or *sensitive attributes*. We note that throughout the paper we will use both terms indistinctly. This variables encode a potential risk of discriminatory information in the population that should not be used by the algorithm. In this framework, two main streams of understanding fairness in machine learning have been considered. The probabilistic notion underlying this division is the independence between distributions. The first one gives rise to the concept of *statistical parity*, which means the independence between the protected attribute and the outcome of the decision rule. This concept is quantified using the Disparate Impact index, which is described for instance in Feldman et al. [2015]. This notion was firstly considered as a tool for quantifying discrimination as the so-called $4/5^{th}$-rule by the State of California Fair Employment Practice Commission (FEPC) in 1971. For more details on the origin and first applications of this index we refer to Biddle [2006]. The second one proposes the *equality of odds*, which considers the independence between the protected attribute and the output prediction, conditionally to the true output value. In other words, it quantifies the independence between the error of the algorithm and the protected variable. Hence, in practice, it compares the error rates of the algorithmic decisions between the different groups of the population. This second point of view has been originally proposed for recidivism of defendants in Flores et al. [2016]. Many others criteria (see for instance in Berk et al. [2018] for a review) have been proposed leading sometimes to incompatible formulations as stated in Chouldechova [2017]. Note finally that the notion of fairness is closely related to the notion of privacy as pointed out in Dwork et al. [2012].

In this paper, our goal is to present some comprehensive statistical results on fairness in machine learning studying the statistical parity criterion through the analysis of the example given in the *Adult Income* dataset. This public dataset is available on the UCI Machine Learning Repository[1] and it consists in forecasting a binary variable (low or high income) which corresponds to an income lower or higher than 50k$ a year. This decision could be potentially used to evaluate the credit risk of loan applicants, making this dataset particularly popular in the machine learning community. It is considered here as potentially sensitive to a discrimination with respect to the *Gender* and *Ethnic origin* variables. The co-variables used in the prediction as well as the true outcome are available in the dataset, hence supervised machine learning algorithms will be used.

Section 2.2 describes this dataset. It specifically highlights the existing unbalance between the income prediction and the *Gender* and *Ethnic origin* sensitive variables. We note that a preprocessing step is needed in order to prepare the data for further analyses and the performed modifications are detailed in the Appendix 2.7.1.1. In Section 2.3, we then explain the statistical framework for the fairness problem, by particularly focusing on the binary classification setting. We follow the approach of the *statistical parity* to quantify the fairness and we thus present the Disparate Impact as our preferred index for measuring the bias. Note that the bias is present in this dataset, so the machine learning decision rules learned in this paper will be trained by using a biased dataset. Although, many criteria have been described in the fair learning literature, they are often used as a score without statistical control. In the cases where test procedures or confidence bounds are provided, they are obtained using a resampling scheme to get standard-

[1] `https://archive.ics.uci.edu/ml/datasets/adult`

ized Gaussian confidence intervals under a Gaussian assumption which does not correspond to the distribution of the observations. In this work, we promote the use of confidence intervals to control the risk of false discriminatory assessment. We then show in the Appendix 2.7.2 the exact asymptotic distribution of the estimates of different fairness criteria obtained through the classical approach of the Delta method described in Van der Vaart [1998]. Then, Section 2.4 is devoted to present some naive approaches that try to correct the discriminatory behaviour of machine learning algorithms or to test possible discriminations. Finally, Section 2.5 is devoted to studying the efficiency of two easy way to incorporate fairness in machine learning algorithms: building a differentiate algorithm for each class of the population or adapting the decision of a single algorithm in a different way for each subpopulation. We then in Section 2.6 present some conclusions for this work and thus provide a concrete pedagogical example for a better understanding of bias issues and fairness treatment in machine learning. Proofs and more technical details are presented in the Appendix. Relevant code in Python to preprocess the Adult Income dataset and reproduce all the analysis and figures presented in this paper are available at the link `https://github.com/XAI-ANITI/StoryOfBias/blob/master/StoryOfBias.ipynb`. We also provide the French version of this Python notebook at `https://github.com/wikistat/Fair-ML-4-Ethical-AI/blob/master/AdultCensus/AdultCensus-R-biasDetection.ipynb`.

2.2 Machine learning algorithms for the attribution of bank loans

One of the applications for which machine learning algorithms have already become firmly established is credit scoring. In order to minimize its risks, the banking industry uses machine learning models to detect the clients who are likely to deal with a credit loan. The FICO score in the US or the SCHUFA score in Germany are examples of these algorithmically determined credit rating scores, as well as those used by a number of Fintech startups, who are also basing their loan decisions entirely on algorithmic models [Hurley and Adebayo, 2016][2]. Yet, credit rating systems have been criticized as opaque and biased in Pasquale [2015], Rothmann et al. [2014] or Hurley and Adebayo [2016].

In this paper, we use the Adult Income dataset as a realistic material to reproduce this kind of analyses for credit risk assessment. This dataset was built by using a database containing the results of a census made in the United States in 1994. It has been largely used among the fair learning community as a suitable benchmark to compare the performance of different machine learning methods. It contains information from about 48 thousands of individuals, each of them being described by 14 variables as detailed in Table 2.2. This dataset is often used to predict the binary variable *Anual Income higher or not than* $50k$\$. Such forecast does not convey any discrimination itself, but it illustrates what can be done in the banking or insurance industry since the machine learning procedures are similar to those made by banks to evaluate the credit risk of their clients. The fact that the true value of the target variable is known, in contrast to the majority of the datasets available in the literature (e.g. the German Credit Data), as well as the value of potential protected attributes such as the ethnic origin or the gender, makes this dataset one of the most widely used to compare the properties of the fair learning algorithms. In this paper, we will then compare supervised machine learning methods on this dataset. A graphic representation of the distribution of each feature can be found in `https://www.valentinmihov.com/2015/04/17/adult-income-data-set/`. This representation gives a good overview of what this dataset contains. It also makes clear that it has to be preprocessed before its analysis using black-box machine learning algorithms. In this work, we have

[2] See, e.g., `https://www.kreditech.com/`.

deleted missing data, errors or inconsistencies. We also have merged highly dispersed categories and eliminated strong redundancies between certain variables (see details in Supplementary material 2.7.1.1). In Figure 2.1, we represent the dataset after our pre-treatments, and show the number of occurrences for each categorical variable as well as the histograms for each continuous variable.

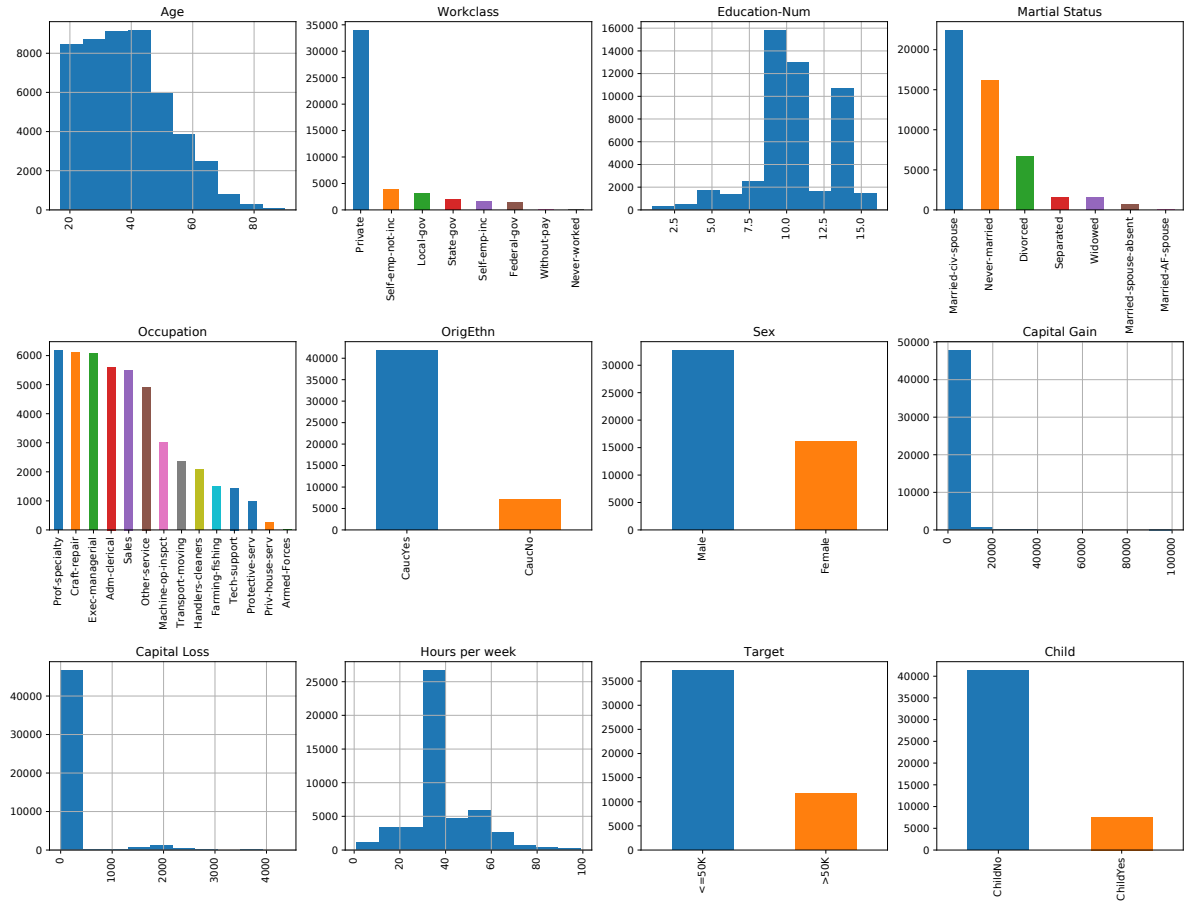


Figure 2.1 – Adult Income dataset after pre-processing phase

2.2.1 Unbalanced Learning Sample

After pre-processing the dataset, standard preliminary exploratory analyses first show that the dataset obviously suffers from an unbalanced repartition of low and high incomes with respect to two variables: *Gender* (male or female) and *Ethnic origin* (caucasian or non-caucasian). These variables therefore seem to be potentially sensitive variables in our data. Figure 2.2 shows this unbalanced repartition of incomes with respect to these variables. It is of high importance to be aware of such unbalanced repartitions in reference datasets since a bank willing to use an automatic algorithm to predict which clients should have successful loan applications could be tempted to train the decision rules on such unbalanced data. This fact is at the heart of our work and we question its effect on further predictions on other data. What information will be learnt from such unbalanced data: a fair relationship between the variables and the true income that will enable socially reasonable forecasts; or biased relations in the repartition of the income with respect to the sensitive variables? We explore this question in the following section.

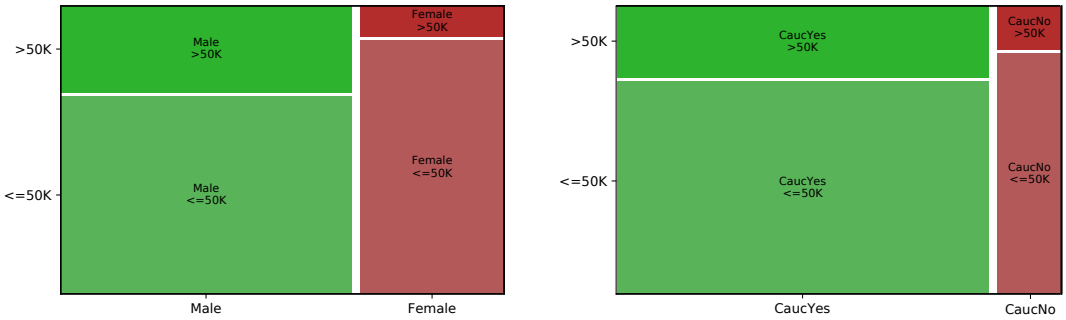


Figure 2.2 – Enbalancement of the reference decisions in the *Adult Income* dataset with respect to the *Gender* and *Ethnic origin* variables.

2.2.2 Machine Learning Algorithms to forecast income

We study now the performance of four categories of supervised learning models: logistic regression [Cramer, 2002], decision trees [Mitchell et al., 1997], gradient boosting [Sutton, 2005], and Neural Network. We used the *Scikit-learn* implementations of the Logistic Regression (LR) and Decision Trees (DT), and the *lightGBM* implementation of the Gradient Boosting (GB) algorithm. The Neural Network (NN) was finally coded using *PyTorch* and contains four fully connected layers with Rectified Linear Units (ReLU) activation functions.

In order to analyze categorical features using these models, the binary categorical variables were encoded using zeros and ones. The categorical variables with more than two classes were also transformed into one-hot vectors, *i.e.* into vectors where only one element is non-zero (or hot). We specifically encoded the target variable by the values $Y = 0$ for an income below $50K\$$, and $Y = 1$ for an income above $50K\$$. We used a 10-fold cross-validation approach in order to assess the robustness of our results. The average accuracy as well as its true positive (TP) and true negative (TN) rates were finally measured for each trained model. Figure 2.3 summarizes these results.

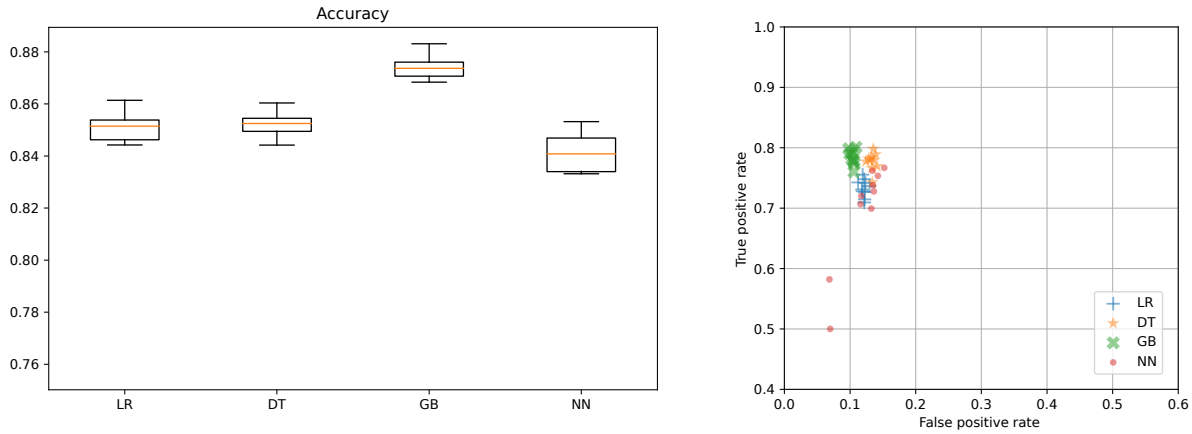


Figure 2.3 – Prediction accuracies, true positive rates and true negative rates obtained by using no specific treatment. Logistic Regression (LR), Decision Tree (DT), Gradient Boosting (GB) and Neural Network (NN) models were tested with 10-folds cross validation on the *Adult Income* dataset.

We can observe in Fig. 2.3 that the best average results are obtained by using Gradient Boosting. More interestingly, we can also remark that the prediction obtained using all mod-

els for $Y = 0$ (represented by the true negative rates) are clearly more accurate than those obtained for $Y = 1$ (represented by the true positive rates), which contains about 24% of the observations. All tested models then make more mistakes on average for the observations which should have a successful prediction than a negative one. Note that the tested neural network is outperformed by other methods in these tests in terms of prediction accuracy. Although we used default parametrizations for the Logistic Regression model as well as the Gradient Boosting model, and we simply tuned the decision tree to have a maximum depth of 5 nodes, we tested different parametrizations of the Neural Network model (number of epochs, mini-batch sizes, optimization strategies) and kept the best performing one. It therefore appears that the neural network model we tested was clearly not adapted to the *Adult Income* dataset.

Hence we have built and compare several algorithms ranging from completely interpretable models to black box models involving optimization of several parameters. Note that we could have used the popular Random Forest algorithm that could lead to equivalent but we privilegiated boosting models whose implementation is easier using Python.

2.3 Measuring the Bias with Disparate Impact

2.3.1 Notations

Among the criteria proposed in the literature to reveal the presence of a bias in a dataset or in automatic decisions (see *e.g.* Hardt et al. [2016] for a recent review), we focus in this paper on the so-called *statistical parity*. This criterion deals with the differences in reference decisions or the outcome of decision rules with respect to a sensitive attribute. Note that we only consider the binary classification problem with a single sensitive attribute for the sake of simplicity, although we could consider other tasks (*e.g.* regression) or multiple sensitive attributes (see Hébert-Johnson et al. [2018] or Kearns et al. [2018]). Here is a summary of the notations we use:

- Y is the variable to be predicted. We consider here binary variables where $Y = 1$ is a positive decision (here a high income) while $Y = 0$ is a negative decision (here a low income);
- $g(X) = \hat{Y}$ is the prediction given by the algorithm. As for Y, this is a binary variable interpreted such that $\hat{Y} = 0$ or $\hat{Y} = 1$ means a negative or a positive decision, respectively. Note that most machine learning algorithms output continuous scores or probabilities. We consider in this case that this output is already thresholded.
- S is the variable which splits the observations into groups for which the decision rules may lead to discriminative outputs. From a legal or a moral point of view, S is a sensitive variable that should not influence the decisions, but could lead to discriminative decisions. We consider hereafter that $S = 0$ represents the minority that could be discriminated, while $S = 1$ represents the majority. We specifically focus here on estimating the disproportionate effect with respect to two sensitive variables: the gender (male vs. female) and the ethnic origin (caucasian vs. non-caucasian).

Statistical parity is often quantified in the fair learning literature using the so-called *disparate impact* (DI). The notion of DI has been introduced in the US legislation in 1971[3]. It measures the existing bias in a dataset as

$$DI(Y, S) = \frac{\mathbb{P}(Y = 1|S = 0)}{\mathbb{P}(Y = 1|S = 1)}, \tag{2.3.1}$$

[3]https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

Table 2.1 – Bias measured in the original dataset

Protected attribute	DI	CI
Gender	0.3597	$[0.3428, 0.3765]$
Ethnic origin	0.6006	$[0.5662, 0.6350]$

and can be empirically estimated as

$$\frac{n_{10}}{(n_{00}+n_{10})} \Big/ \frac{n_{11}}{(n_{01}+n_{11})}, \tag{2.3.2}$$

where n_{ij} is number of observations such that $Y = i$ and $S = j$. The smaller this index, the stronger the discrimination over the minority group. Note first that this index supposes that $\mathbb{P}(Y = 1|S = 0) < \mathbb{P}(Y = 1|S = 1)$ since S is defined as the group which can be discriminated with respect to the output Y. It is also important to remark that this estimation may be unstable due to the unbalanced amount of observations in the groups $S = 0$ and $S = 1$ and the inherent noise existing in all data. We then propose to estimate a confidence interval around the *disparate impact* in order to provide statistical guarantees of this score, as detailed in the Supplementary material 2.7.2. These confidence intervals will be used later in this section to quantify how reliable are two disparate impacts computed on our dataset. This fairness criterion can be extended to the outcome of an algorithm by replacing in Eq. (2.3.1) the true variable Y by $g(X) = \hat{Y}$, that is

$$DI(g, X, S) = \frac{\mathbb{P}(g(X, S) = 1|S = 0)}{\mathbb{P}(g(X, S) = 1|S = 1)}. \tag{2.3.3}$$

This measures the risk of discrimination when using the decision rules encoded in g on data following the same distribution as in the test set. Hence, in Gordaliza et al. [2019] a classifier g is said not to have a *disparate impact* at level $\tau \in (0, 1]$ when $DI(g, X, S) > \tau$. Note that the notion of DI defined Eq. (2.3.1) was first introduced as the $4/5^{th}$-rule by the State of California Fair Employment Practice Commission (FEPC) in 1971. Since then, the threshold $\tau_0 = 0.8$ was chosen in different trials as a legal score to judge whether the discriminations committed by an algorithm are acceptable or not (see e.g. Feldman et al. [2015] Zafar et al. [2017a], or Mercat-Bruns [2016]).

2.3.2 Measures of disparate impacts

The *disparate impact* $DI(g, X, S)$ should be obviously close to 1 to claim that g makes fair decisions. A more subtle, though critical, remark is that it should at least not be smaller than the general disparate impact $DI(Y, S)$. This would indeed mean that the decision rules g reinforce the discriminations compared with the reference data on which it was trained. We will then measure hereafter the disparate impacts $DI(Y, S)$ and $DI(g, X, S)$ obtained on our dataset.

In Table 2.1, we have quantified confidence intervals for the bias already present in the original dataset using Eq. (2.3.1) with the sensitive attributes *Gender* and *Ethnic origin*. They were computed using the method of Appendix 2.7.2 and represent the range of values the computed disparate impacts can have with a 95% confidence (subject to standard and reasonable hypotheses on the data). Here the DI computed on the *Gender* variable then appears as very robust and the one computed on the *Ethnic origin* variable is relatively robust. It is clear from this table that both considered sensitive attributes generate discriminations. These discriminations are also more severe for the *Gender* variable than for the *Ethnic origin* variable.

We have then measured the disparate impacts Eq. (3.3.6) obtained using the predictions made by the four models in the 10-folds cross-validation of Section 2.2.2. These disparate impacts are presented in Fig. 2.4. We can see that, except for the decision tree with the *Ethnic origin* variable, the algorithms have smaller disparate impact than for the true variable. The impact is additionally clearly worsened with the *Gender* variable using all trained predictors. These predictors therefore reinforced the discriminations in all cases by enhancing the bias present in the training sample. Observing the true positive and true negative rates of Fig. 2.4, which distinguish the groups $S = 0$ and $S = 1$ is particularly interesting here to understand this effect more deeply. As already mentioned Section 2.2.2, the true negative (TN) rates are generally higher than the true positive (TP) rates. It can be seen Fig. 2.4 that this phenomenon is clearly stronger in the subplot representing the TP and TN for $S = 0$ than the one representing them for $S = 1$, so false predictions are more favorable to the group $S = 1$ than the group $S = 0$. This explains why the disparate impacts of the predictions are higher than those of the original data (boxplots *Ref* in Fig. 2.4). Note that these measures are directly related to the notions of equality of odds and opportunity as discussed in Hardt et al. [2016]. The machine learning models we used in our experiments were then shown as unfair on this dataset, in the sense that discrimination is reinforced.

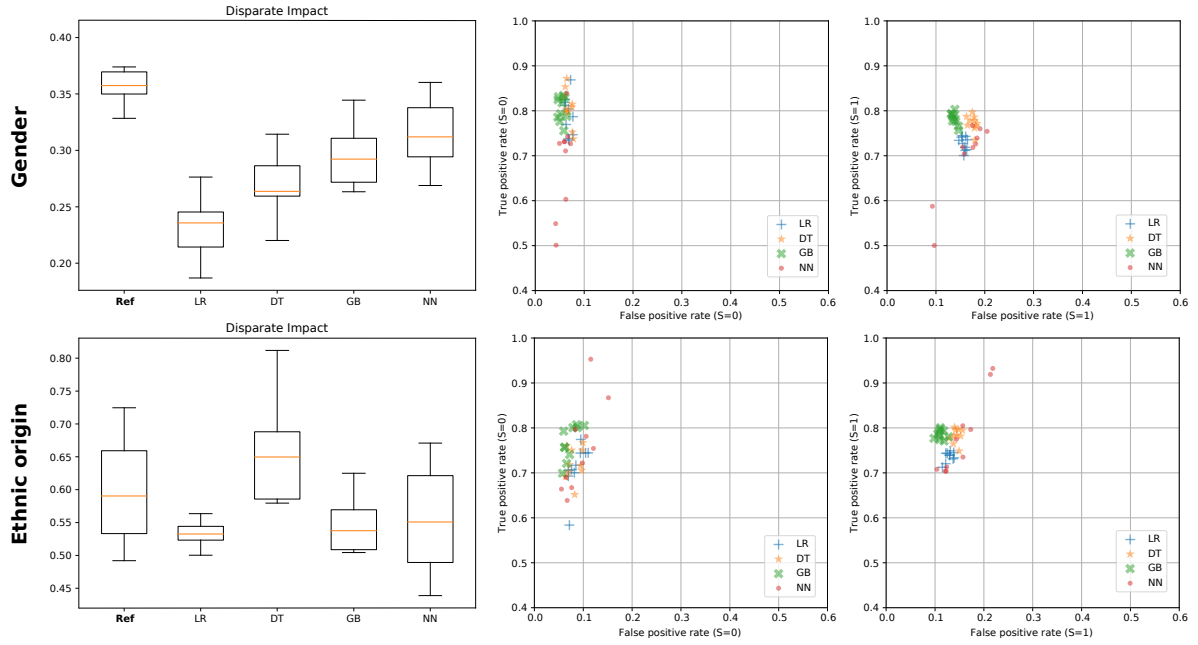


Figure 2.4 – Bias measured in the outputs of the tested machine learning models (LR, DT, GB, NN) using the 10-folds cross validation. The disparate impacts of the reference decisions are represented by the boxplot *Ref* to make clear that the unfairness is almost always re-inforced in our tests by automatic decisions. These is also a good balance between the true and the false positive decisions when the results are close to the dashed blue line. **(Top)** *Gender* is the sensitive variable. **(Bottom)** *Ethnic origin* is the sensitive variable.

As pointed out in Friedler et al. [2019], there may have a strong variability when computing the disparate impact of different subsamples of the data. Hence, we additionally propose in this paper an exact Central Limit Theorem to overcome this effect. The confidence intervals we obtain prove their stability when confronted to bootstrap replications and for this therefore cross-validated our results using 10 replications of different learning and test samples on the

three algorithms. The construction of these confidence intervals are postponed to Section 2.7.2 while comparison with bootstrap procedures are detailed in Section 7.3 of the Appendix. In order to conveniently compare the bias in the predictions with the one in the original data, we show on the left the bias measured in the data. We can see that these boxplots are coherent with the results of Table 2.1 and Figure 2.4, and again show that the discrimination was reinforced by the machine learning models in this test.

In all generality, we conclude here that one has to be careful when training decision rules. They can indeed worsen existing discriminations in the original database. We also remark that the majority of works using the Disparate Impact as a measure of fairness rely only on this score as a numerical value with no estimation of how reliable it is. This motivated the definition of our confidence intervals strategy in Appendix 2.7.2, which was shown to be realistic in our experiments when comparing the *Ref* boxplots of Figure 2.4 with the confidence intervales of Tables 2.1. Note that we will only focus in the rest of the paper on the protected variable *Gender* since it was shown in Section 2.3 to be clearly the variable leading to discrimination for all tested machine learning models. We will also only test the Logistic Regression (LR) and Decision Tree (DT) as they are highly interpretable, plus the Gradient Boosting (GB) model which was shown to be the best performing one on the *Adult Census* dataset.

2.4 A quantitative evaluation of GDPR recommendations against algorithm discrimination

Once the presence of bias is detected, the goal of machine learning becomes to reduce its impact without hampering the efficiency of the algorithm. Actually, the predictions made by the algorithm should remain sufficiently accurate to make the machine learning model relevant in Artificial Intelligence applications. For instance, the decisions $\hat{Y}$ made by a well balanced coin when playing *head or tail* are absolutely fair, as they are independent of any possible sensitive variable S. However, they also do not take into account any other input information X, making them pointless in practice. Reducing the bias of a machine learning model g therefore ideally consists in getting rid of the influence of S in all input data (X, S) while preserving the relevant information to predict the true outputs Y. We will see below that this is not that obvious, even in our simple example.

It is first interesting to remark that the problem cannot be solved by simply having a balanced amount of observations with $S = 0$ and $S = 1$. We indeed reproduced the experimental protocol of Section 2.3.2 with 16,192 randomly chosen observations representing males (instead of 32,650), so that the decision rules were trained in average with as many males as females. As shown in Fig. 2.5, the trends of the results turned out to be very similar to those obtained in Fig. 2.4-*(Gender)*.

We specifically study in section the effect of complying to the European regulations. From a legal point of view, the GDPR's recommendation indeed consists in not using the sensitive variable in machine learning algorithms. Hence, we simply remove here S from the database in subsection 2.4.1, and we consider in subsection 2.4.2 one of the most common legal proof for discrimination called the *testing method*. It consists in considering the response for the same individual but with a different sensitive variable. We will study whether this procedure enables to detect the group discrimination coming from the decisions of an algorithm.

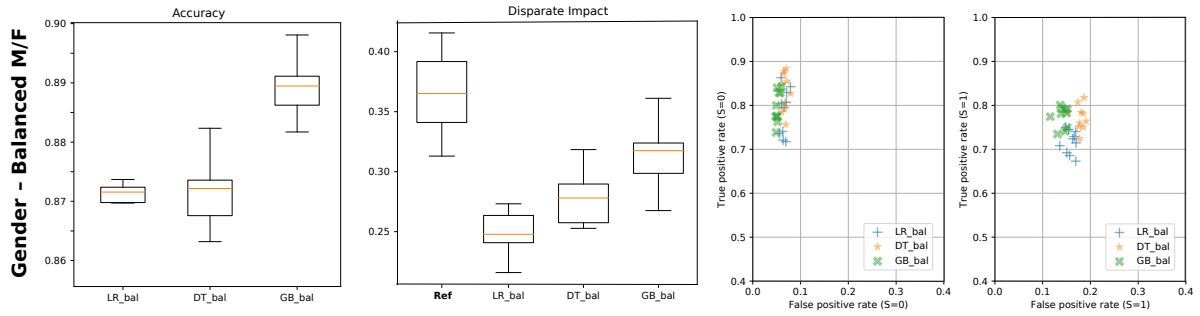


Figure 2.5 – Bias measured in the outputs of the LR, DT and GB machine learning models using the same experimental protocol as in Section 2.3.2 (see specifically Fig. 2.4-*(Gender)*), except that we used the same amount of males ($S = 1$) and females ($S = 0$) in the dataset.

2.4.1 What if the sensitive variable is removed?

The most obvious idea to remove the influence of a sensitive variable S is to remove it from the data, so we cannot use it when training the decision rules and then obviously when making new decisions. Note that this solution is recommended by GPDR regulations. To test the pertinence of this solution, we considered the algorithms analyzed in Sections 2.2 and 2.3 and then used them without using the *Gender* variable. As in Section 2.3, a 10-fold cross-validation approach was used to assess the robustness of our results.

As shown Figure 2.6-(top), the disparate impacts as well as the model accuracies remained almost unchanged when removing the *Gender* variable from the input data. Anonymizing database by removing a variable therefore had very little effect on the discrimination that is induced by the use of an automated decision algorithm. This is very likely to be explained by the fact that a machine learning algorithm uses all possible information conveyed by the variables. In particular, if the sensitive variable (here the *Gender* variable) is strongly correlated to other variables, then the algorithm learns and reconstruct automatically the sensitive variable from the other variables. Hence we can deduce that social determinism is stronger than the presence of the sensitive variable here, so the classification algorithms were not impacted by the removal of this variable.

Obtaining fairness is a far more complicated task than this simple trick. It is at the heart of modern research on fair learning. More complex fairness mathematical methods to reduce disparate treatment are discussed for instance in Kleinberg et al. [2016] or in Gordaliza et al. [2019].

2.4.2 From Testing for bias detection to unfair prediction

Testing procedures are often used as a legal proof for discrimination. For an individual prediction, such procedures consist in first creating an artificial individual which shares the same characteristics of a chosen individual that suspects a disparate treatment and discrimination, but has a different protected variable. Then it amounts to testing whether this artificial individual has the same prediction as the original one. If the predictions differ, then this conclusion can serve as a legal proof for discrimination.

These procedures have existed for a long time (since their introduction in 1939 [4]) , and since 2006 when the French justice has taken them as a proof of biased treatment, although the

[4] `https://fr.wikipedia.org/wiki/Test_de_discrimination`

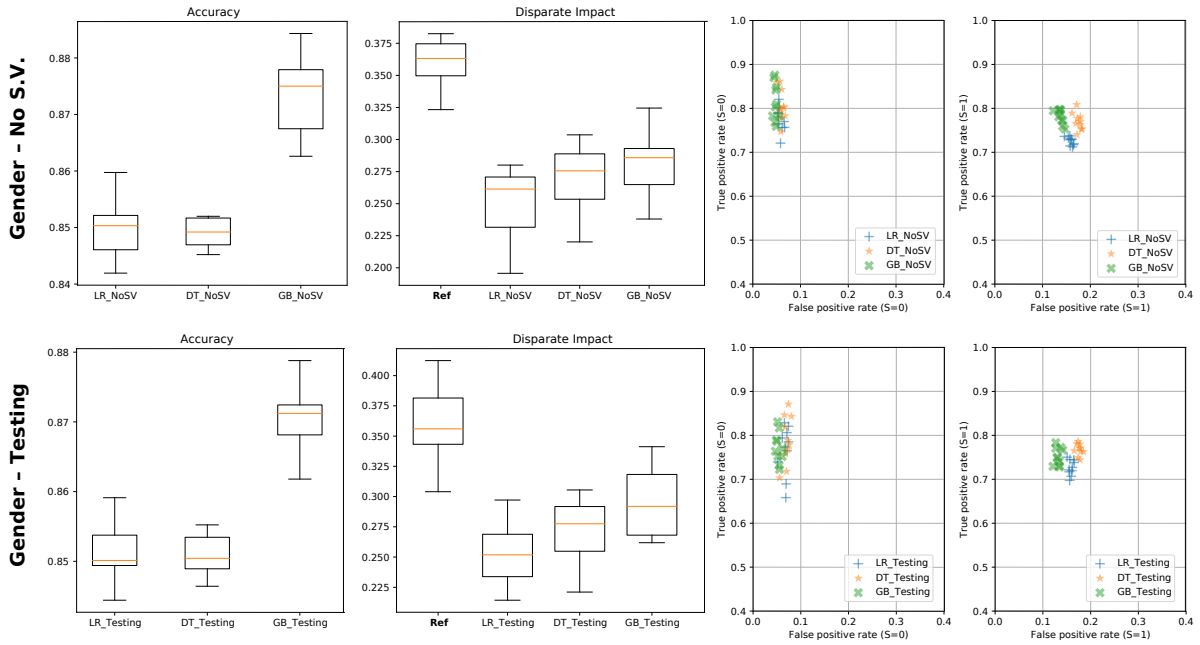


Figure 2.6 – Performance of the machine learning models LR, DT and GB when **(top)** removing the *Gender* variable, and **(bottom)** when using a testing procedure.

testing process itself has been qualified as unfair[5]. Furthermore, this technique has been generalized by sociologists ans economists (see for instance Riach and Rich [2002] for a description of such method) to statistically measure group discrimination in housing and labour market by conducting carefully controlled field experiments.

This testing procedure considered as a discrimination test is nowadays a commonly used method in France to assess fairness for sociological studies of *Observatoire des discriminations*[6] and *laboratoire TEPP* as pointed out in L'Horty et al., or governemental studies DARES[7] of French Ministry of Work ISM Corum [8]. Some industries are labeled using such test. An audit quality of recruiting methods is proposed while *Novethic*[9] proposes ethic formations.

Testing is efficient to detect human discrimination specially in labour market but hiring tech is producing more and more softwares or web platforms performing predictive recruitment as in Raghavan et al. [2020]. Does testing remain valid in front of machine learning algorithms? This last strategy is evaluated using the same experimental protocol as in the previous sections. The results of these experiments are shown in Figure 6-(bottom). Testing does not detect any discrimination when the sensitive variable is captured by the other variables.

An algorithmic solution to bypass this testing procedure is given by the following trick. Train a classifier as usual using all available information X, S and then build a *testing compliant* version of it as follows : for an individual, the predicted outcome is assigned as the best decision obtained on the actual individual $f(x, s)$ and a virtual individual with exactly the same characteristics as the original one, except for the protected variable s which has the opposite label $s^{'}$ (*e.g.* the *Gender* variable is *Male* instead of *Female*), namely $f(x, s^{'})$. Note that in case of multi-class

[5] https://www.juritravail.com/discrimination-physique/embauche/ph-alternative-A-1.html

[6] https://www.observatoiredesdiscriminations.fr/testing

[7] https://dares.travail-emploi.gouv.fr/dares-etudes-et-statistiques/etudes-et-syntheses/dares-analyses-dares-indicateurs-dares-resultats/testing

[8] http://www.ismcorum.org/

[9] https://www.novethic.fr/lexique/detail/testing.html

labels, the outcome should be the most favourable decision for all possible labels. This classifier is fair by design in the sense that no matter their gender, the *testing procedure* can not detect a change in the individual prediction.

Nevertheless, this trick against testing cannot cheat usual evaluation of discrimination by using a disparate impact measure which is usual in the USA by measuring the impact on real and not fictitious recruitment. This is the reason why hiring tech companies add some facilities (Raghavan et al. [2020]) to mitigate ethnic bias of algorithmic hiring for avoiding an enterprise juridical complications. The evaluation of this strategy is evaluated using the same experimental protocol as in the previous sections and these are shown in Figure 2.6-(bottom).

As expected for previous results, this method has little impact on the classification errors and the disparate impacts. This emphasises the conclusion of Section 2.4.1 claiming that the *Gender* variable is captured by other variables. Removing the effect of a sensitive variable can therefore require more advanced treatments than those described above.

2.5 Differential treatment for fair decision rules

2.5.1 Strategies

As we have seen previously, bias may induce discrimination of an automatic decision rule. Although many complex methods have been developed to tackle this problem, we investigate in this section the effects of two easy and maybe naive modifications of machine learning algorithms. We present in this section the effect of two alternative strategies to build fair classifiers. They have in common the idea of considering different treatments according to each group $S = \{0, 1\}$. These strategies are the following :

1. **Building a different classifier for each class of the sensitive variable:** This strategy consists in training the same prediction model with different parameters for each class of the sensitive variable. We denote *separate treatment* this strategy.

2. **Using a specific threshold for each class of the sensitive variable:** Here, a single classifier is trained for all data to produce a score. The binary prediction is however get using a specific threshold for each sub-group $S = 0$ or $S = 1$. Note that when the score is obtained by estimating the conditional distribution $\eta(x) = P(Y = 1|X = x)$ then the threshold used is often 0.5. Here this threshold is made S-dependent and is adapted to avoid any possible discrimination. In practice, we keep a threshold of 0.5 for the observations in the group $S = 1$ but we adapt the corresponding threshold for the observations in the group $S = 0$. In our tests, we automatically set this threshold on the training set so that the disparate impact is close to 0.8 in the cases where it was originally lower to this this socially accepted threshold. The classifier and the potentially adapted threshold are then used for further predictions. This corresponds in a certain way to favour the minority class by changing equality to equity. We denote this strategy as *positive discrimination* since this procedure corresponds to this purpose.

2.5.2 Results obtained using the *Separate Treatment* strategy

Splitting the model parameters into parameters adapted to each group reduces the bias of the predictions when compared to the initial model, but it does not remove it. As we can see in Figure 2.7-*(top)*, where the notations are analogous to those in the above figures, it improved the disparate impact in all cases for relatively stable prediction accuracies. Note that the improvements are more spectacular for the basic Logistic Regression and Decision Tree

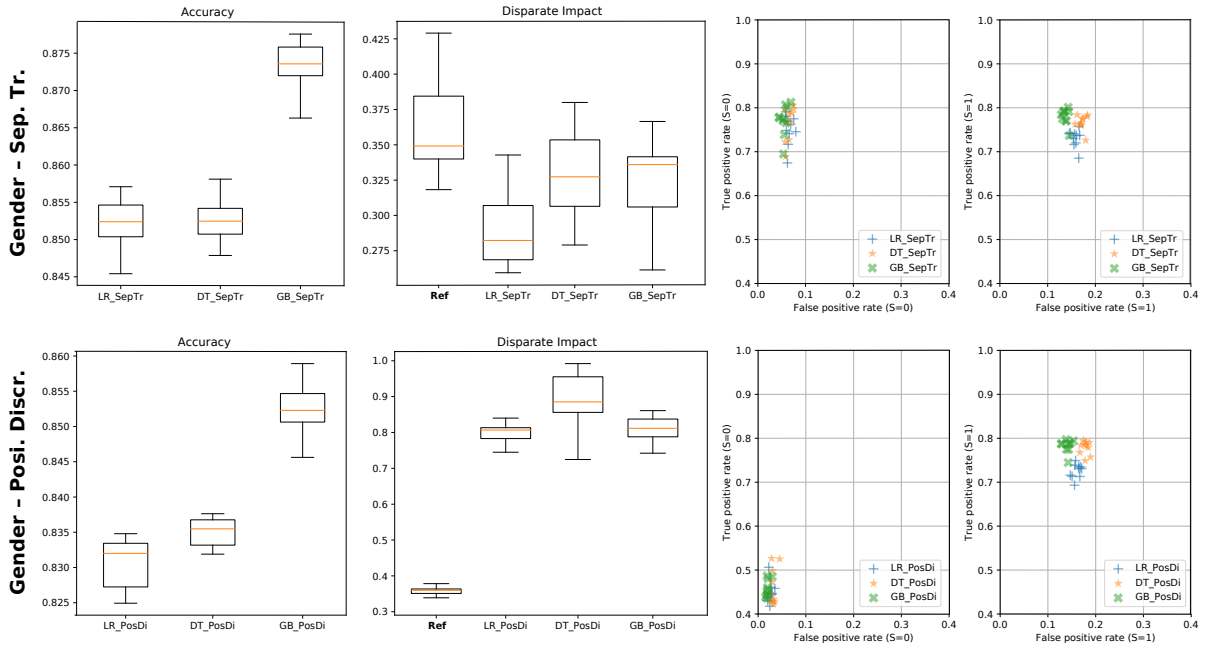


Figure 2.7 – Performance of the machine learning models LR, DT and GB when **(top)** using a *Separate Treatment* for the groups $S = 0$ and $S = 1$, and **(bottom)** when using a *Positive Discrimination* strategy for the groups $S = 0$.

models than for the Gradient Boosting model. This last model is indeed particularly efficient to capture fine high order relations between the variables, which gives less influence to the strong non-linearity generated when splitting the machine learning model into two class-specific models. Hence building different models reduces but does not solve the problem, the level of discrimination in the decisions being only slightly closer to the level of bias in the initial dataset.

2.5.3 Results obtained using the *Positive Discrimination* strategy

Results obtained using the *positive discrimination* strategy are shown in Figure 2.7-*(bottom)*. They clearly emphasize the spectacular effect of this strategy on the disparate impacts, which can be controlled by the data scientist. By adjusting the threshold, it is possible to adjust the levels of discriminations in the dataset, as in this example where the socially acceptable level of 0.8 can be reached. In this case we see a decrease in the performance of the classifier, but yet being reasonable.

These results should however be tempered for a main reason. Although the average error receives little changes, the number of false positive cases of women is clearly increased when introducing positive discrimination. In our tests more than half of the predictions that should have been false in the group $S = 0$ are even true. These false positive decisions have a limited impact on the average prediction accuracy as they where obtained in the group $S = 0$ which has less observations than $S = 1$ and that there are clearly less true predictions with $Y = 1$ than $Y = 0$. Yet false positive errors are considered as the most important error type and thus this increase may be very harmful for the decision maker. On a legal point of view, this procedure may be judged as unfair or rises political issues that are far beyond the scope of this paper.

2.6 Conclusions

In this paper, we provided a case-study of the use of machine learning technics for the prediction of the well-known *Adult Income* dataset. We focused on a specific fairness criterion, the statistical parity, which is measured through the Disparate Impact. This metric quantifies the difference of the behaviour of a classification rule applied for two subgroups of the population, the minority and the majority. Fairness is achieved when the algorithm behaves in the same way for both groups, hence when the sensitive variable does not play a significant role in the prediction. Main results are summarized in Figure 2.8.

In particular, we convey the following take-home messages: **(1)** Bias in the training data may lead to machine learning algorithms taking unfair decisions, but not always. While there is a clear increase of bias using the tested machine learning algorithms with respect to the *Gender* variable, the *Ethnic Origin* does not lead to a severe bias. **(2)** As always in Statistics, computing a mere measure is not enough but confidence intervals are needed to determine the variability of such indexes. Hence, we proposed an ad-hoc construction of confidence intervals for the Disparate Impact. **(3)** Standard regulations that promote either the removal of the sensitive variable or the use of testing technics appeared as irrelevant when dealing with fairness of machine learning algorithms.

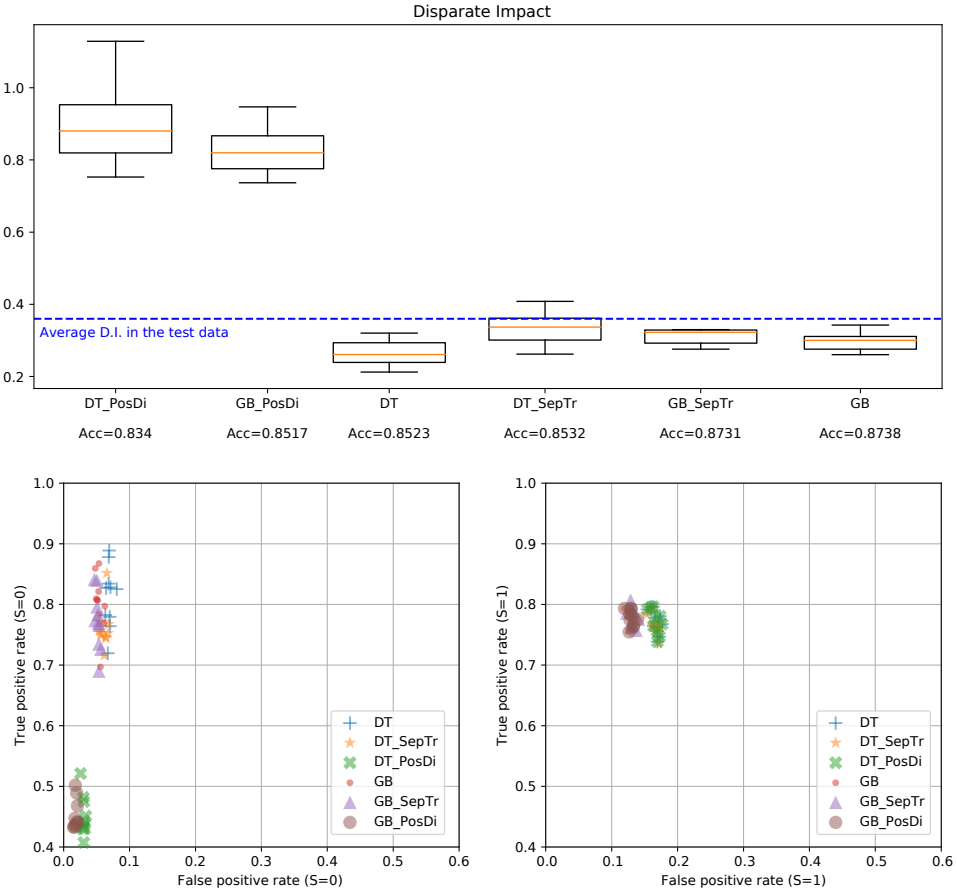


Figure 2.8 – Summary of the main results: The best performing algorithms of Sections 2.3 and 2.5 are compared here. **(top)** Boxplots of the disparate impacts from the least accurate method on the left, to the most accurate method on the right, and **(bottom)** corresponding true positive and true negative rates in the groups $S = 0$ and $S = 1$.

Note also that different notions of fairness (local and global) are at stake here. We first point out that testing methods focus on individual fairness while statistical methods such as the Disparate Impact Analysis tackle the issue of group fairness. These two notions if related to the similar notion of discrimination with respect to an algorithmic decision are yet different. In this work, we showed that an algorithm can be designed to be individually fair while still presenting a strong discrimination with respect to the minority group. This is mainly due to the fact that testing methods are unable to detect the discrimination hidden in the algorithmic decisions that are due to the training on an unbalanced sample. Testing methods detect discrimination if individuals with the same characteristics but different sensitive variables are treated in a different way. This corresponds to trying to find counterfactual explanation to an individual with a different sensitive variable. This notion of counterfactual explanations to detect unfairness has been developed in Kusner et al. [2017]. Yet the testing method fails in finding a counterfactual individual since it is not enough to change only the sensitive variable but a good candidate should be the closest individual with a different sensitive variable but with the variables that evolve depending on S. For this, following some recent work on fairness with optimal transport theory as in Gordaliza et al. [2019] developing an idea from Feldman et al. [2015], some authors propose a new way of testing discrimination by computing such new counterfactual models in Black et al. [2020]. Finally, we tested two a priori naive solutions consisting either in building different models for each group or in choosing different rules for each group. Only the latter that can be considered as *positive discrimination* proves helpful in obtaining a fair classification. Note that if some errors are increased (false positive rate), this method has a good generalization error. Yet in other cases, the loss of efficiency could be greater and this method may lead to unfair treatment.

This data set has been extensively studied in the literature on fairness in machine learning and we are well aware of the numerous solutions that have been proposed to solve this issue. Even with standard methods, it is possible for a data scientist, when confronted to fairness in machine learning, to design algorithms that have very different behaviors and yet achieving a good classification error rate. Some algorithms hamper discrimination in the society while others just maintain its level, and some others correct this discrimination and provide gender equity. It is worth noting that the most explainable algorithms, such as the logistic regression, do not protect from discrimination. On the contrary, the capture of gender bias is inmediate due to its simplicity, while more complex algorithms might be more protected from this spurious correlation or, since the variable is discrete, better said spurious dependency.

The choice of a model should not be driven only by its performance with respect to a generalization error but should also be explainable in terms of bias propagation. For this, measures of fairness should be included in the evaluation of the model. In this work, we only considered statistical parity type fairness but many other definitions are available, without any consensus on the better choice for such a definition neither from a mathematical or a legal point of view. A strong research effort in data science is hence the key for a better use of Artificial Intelligence type algorithms. This will allow data scientists to describe precisely the algorithmic designing process, as well as their behaviour, in terms of precision and propagation of bias.

In closing, note that biases are what enables machine learning algorithms to work and helpfulness of complex algorithms is due to their ability to find hidden bias and correlations in very large data sets. Hence bias removal should be handled with care because one part of this information is crucial, while the other is harmful. Therefore, explainability should not be understood in terms of explainability of the whole algorithm, but maybe one line of future research in machine learning should focus on explainability of the inner bias of an algorithm, or its explainability with respect to some legal regulations.

Table 2.2 – The Adult Income dataset

N°	Label	Possible values
1	`Age`	Real
2	`workClass`	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Withoutpay, Never-worked
3	`fnlwgt`	Real
4	`education`	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
5	`educNum`	integer
6	`mariStat`	Married-civ-spouse, Divorced, Nevermarried, Separated, Widowed, Marriedspouse- absent, Married-AF-spouse
7	`occup`	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transportmoving, Priv-house-serv, Protective-serv, Armed-Forces
8	`relationship`	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
9	`origEthn`	White, Asian-Pac-Islander, Amer-Indian- Eskimo, Other, Black
10	`gender`	Female, Male
11	`capitalGain`	Real
12	`capitalLoss`	Real
13	`hoursWeek`	Real
14	`nativCountry`	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying- US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad and Tobago, Peru, Hong, Holand- Netherlands
15	`income`	$> 50k$, $\leq 50k$

2.7 Appendix to Chapter 2

2.7.1 The Adult Income dataset

2.7.1.1 Data preparation

As discussed in the introduction of Section 2.2, the study has started with a detailed preprocessing of the raw data to give a more clear interpretation to further analyses. First, we noticed that the variable `fnlwgt` (Final sampling weight) has not a very clear meaning so it has been removed. For a complete description of such variable access the link `http://web.cs.wpi.edu/~cs4341/C00/Projects/fnlwgt`. We have also performed a basic and multidimensional exploration (MFCA) to represent the possible sources of bias in the data in `https://github.com/wikistat/Fair-ML-4-Ethical-AI/blob/master/AdultCensus/AdultCensus-R-biasDetection.ipynb`.

This exploration leaded to a deep cleaning of the data set and highlighted difficulties present on certain variables, raising the need to transform some of them before fitting any statistical model. In particular, we have deleted missing data, errors or inconsistencies; grouped together certain highly dispersed categories and eliminated strong redundancies between certain variables. This phase is notoriously different from the strategy followed by Friedler et al. [2019] who analyze

raw data directly. Some of these main changes are listed below:

- Variable 3 `fnlwgt` is removed since it has little significance for this analysis.
- The binary variable `child` is created to indicate the presence or absence of children.
- Variable 8 `relationship` is removed since it is redundant with `gender` and `mariStat`.
- Variable 14 `nativCountry` is removed since it is redundant with variable `origEthn`.
- Variable 9 `origEthn` is transformed into a binary variable: CaucYes vs. CaucNo.
- Varible 4 `education` is removed as redundant with variable `educNum`.
- Additionally clean-up the $< 50K$, $\leq 50K$, $> 50K$ and $\geq 50K$ in variable "Target"

2.7.2 Testing lack of fairness and confidence intervals

Let $\left(X_i, S_i, \hat{Y}_i = g(X_i)\right), i = 1, \ldots, n$, be a random sample of independent and equally distributed variables. Previous criterion can be consistently estimated by their empirical version. Yet the value of the criterion may depend on the data sample. Due to the importance of obtaining an accurate proof of unfairness in a decision rule it is important to obtain confidence intervals in order to control the error of detecting unfairness. In the literature it is often achieved by computing the mean over several sampling of the data. We provide in the following the exact asymptotic behaviors of the estimates in order to build confidence intervals.

Theorem 2.7.1 (Asymptotic behavior of the *disparate impact* estimator) *Set the empirical estimator of DI(g) as*

$$T_n := \frac{\sum_{i=1}^n \mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}\sum_{i=1}^n \mathbb{1}_{S_i=1}}{\sum_{i=1}^n \mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1}\sum_{i=1}^n \mathbb{1}_{S_i=0}}.$$

Then the asymptotic distribution of this quantity is given by

$$\frac{\sqrt{n}}{\sigma}\left(T_n - DI(g, X, S)\right) \xrightarrow{d} N(0,1), \ as \ n \to \infty, \tag{2.7.1}$$

where $\sigma = \sqrt{\nabla\varphi^T\left(\mathbb{E}Z_1\right)\Sigma_4\nabla\varphi\left(\mathbb{E}Z_1\right)}$ *and*

$$\nabla\varphi^T\left(\mathbb{E}Z_1\right) = \left(\frac{\pi_1}{p_1\pi_0}, -\frac{p_0\pi_1}{p_1^2\pi_0}, -\frac{p_0\pi_1}{p_1\pi_0^2}, \frac{p_0}{p_1\pi_0}\right)$$

$$\Sigma_4 = \begin{pmatrix} p_0(1-p_0) & & & \\ -p_0p_1 & p_1(1-p_1) & & \\ \pi_1p_0 & -\pi_0p_1 & \pi_0\pi_1 & \\ -\pi_1p_0 & \pi_0p_1 & -\pi_0\pi_1 & \pi_0\pi_1 \end{pmatrix},$$

where we have denoted $\pi_s = \mathbb{P}(S_1 = s)$ *and* $p_s = \mathbb{P}(g(X_1) = 1, S_i = s), \ s = 0, 1, .$

Proof:

Consider for $i = 1, \ldots, n$, the random vectors

$$Z_i = \begin{pmatrix} \mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0} \\ \mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1} \\ \mathbb{1}_{S_i=0} \\ \mathbb{1}_{S_i=1} \end{pmatrix},$$

where $\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=s} \sim B(\mathbb{P}(g(X_i)=1, S_i=s))$ and $\mathbb{1}_{S_i=s} \sim B(\mathbb{P}(S_i=s)),\ s=0,1,$. Thus, Z_i has expectation

$$\mathbb{E}Z_i = \begin{pmatrix} \mathbb{P}(g(X_i)=1, S_i=0) \\ \mathbb{P}(g(X_i)=1, S_i=1) \\ \mathbb{P}(S_i=0) \\ \mathbb{P}(S_i=1) \end{pmatrix}.$$

The elements of the covariance matrix Σ_4 of Z_i are computed as follows:

$$Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}, \mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1}\right) = \mathbb{E}\left(\mathbb{1}^2_{g(X_i)=1}\mathbb{1}_{S_i=0}\mathbb{1}_{S_i=1}\right) - \mathbb{P}(g(X_i)=1, S_i=0)\mathbb{P}(g(X_i)=1, S_i=1)$$

$$\begin{aligned} Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}, \mathbb{1}_{S_i=0}\right) &= \mathbb{E}\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}^2_{S_i=0}\right) - \mathbb{P}(g(X_i)=1, S_i=0)\mathbb{P}(S_i=0) \\ &= \mathbb{P}(g(X_i)=1)\mathbb{P}(S_i=0) - \mathbb{P}(g(X_i)=1, S_i=0)\mathbb{P}(S_i=0) \\ &= \left[1-\mathbb{P}(S_i=0)\right]\mathbb{P}(g(X_i)=1, S_i=0) \end{aligned}$$

$$Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}, \mathbb{1}_{S_i=1}\right) = \mathbb{E}\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}\mathbb{1}_{S_i=1}\right) - \mathbb{P}(g(X_i)=1, S_i=0)\mathbb{P}(S_i=1)$$

$$Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1}, \mathbb{1}_{S_i=0}\right) = \mathbb{E}\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}\mathbb{1}_{S_i=1}\right) - \mathbb{P}(g(X_i)=1, S_i=1)\mathbb{P}(S_i=0)$$

$$\begin{aligned} Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1}, \mathbb{1}_{S_i=1}\right) &= \mathbb{E}\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}^2_{S_i=1}\right) - \mathbb{P}(S_i=1)\mathbb{P}(g(X_i)=1, S_i=1) \\ &= \mathbb{P}(g(X_i)=1, S_i=1) - \mathbb{P}(S_i=1)\mathbb{P}(g(X_i)=1, S_i=1) \\ &= \mathbb{P}(g(X_i)=1, S_i=1)\left[1-\mathbb{P}(S_i=1)\right] \\ &= \mathbb{P}(g(X_i)=1, S_i=1)\mathbb{P}(S_i=0) \end{aligned}$$

and finally,

$$Cov(\mathbb{1}_{S_i=0}, \mathbb{1}_{S_i=1}) = \mathbb{E}\left(\mathbb{1}_{S_i=0}\mathbb{1}_{S_i=1}\right) - \mathbb{P}(S_i=0)\mathbb{P}(S_i=1) = -\mathbb{P}(S_i=0)\mathbb{P}(S_i=1).$$

From the Central Limit Theorem in dimension 4, we have that

$$\sqrt{n}\left(\bar{Z}_n - \mathbb{E}Z_1\right) \xrightarrow{d} N_4\left(\mathbf{0}, \Sigma_4\right),\ as\ n \to \infty.$$

Now consider the function

$$\begin{array}{rccc} \varphi: & \mathbb{R}^4 & \longrightarrow & \mathbb{R} \\ & (x_1, x_2, x_3, x_4) & \longmapsto & \dfrac{x_1 x_4}{x_2 x_3} \end{array}$$

Applying the Delta-Method (see in Van der Vaart [1998]) for the function φ, we conclude that

$$\sqrt{n}\left(\varphi(\bar{Z}_n) - \varphi(\mathbb{E}Z_1)\right) \xrightarrow{d} \nabla\varphi^T\left(\mathbb{E}Z_1\right) N_4\left(\mathbf{0}, \Sigma_4\right),\ as\ n \to \infty,$$

where $\varphi(\bar{Z}_n) = T_n,\ \varphi(\mathbb{E}Z_1) = DI(g, X, S)$. $\square$

Hence, we can provide a confidence interval when estimating the disparate impact over a data set. Actually $\left(T_n \pm \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}\right)$ is a confidence interval for the parameter $DI(g, X, S)$ asymptotically of level $1-\alpha$.

Previous theorem can be used to test the presence of disparate impact at a given level.

$$H_{0,\beta}:\ DI(g, X, S) \leqslant \beta \quad vs. \quad H_{1,\beta}:\ DI(g, X, S) > \beta \tag{2.7.2}$$

aims at checking if g has Disparate Impact at level β. We want to check wether $DI(g, X, S) \leq \beta$. Under H_0, the inequality $T_n - \beta \leqslant T_n - DI(g, X, S)$ holds, and so

$$\frac{\sqrt{n}}{\sigma}\left(T_n - \beta\right) \leqslant \frac{\sqrt{n}}{\sigma}\left(T_n - DI(g, X, S)\right).$$

Finally, from the inequality above and Eq. (2.7.1), we have that

$$\mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}\left(T_n - \beta\right) < Z_{1-\alpha}\right) \geqslant \mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}\left(T_n - DI(g, X, S)\right) < Z_{1-\alpha}\right) \longrightarrow 1 - \alpha,$$

as $n \to \infty$ and, equivalently,

$$\mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}\left(T_n - \beta\right) \geqslant Z_{1-\alpha}\right) \leqslant \mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}\left(T_n - DI(g, X, S)\right) \geqslant Z_{1-\alpha}\right) \longrightarrow \alpha,$$

as $n \to \infty$, where $Z_{1-\alpha}$ is the $(1-\alpha)$-quantile of $N(0,1)$. In conclusion, the test rejects H_0 at level α when

$$\mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}\left(T_n - \beta\right) \geqslant Z_{1-\alpha}\right) \geqslant \alpha.$$

When dealing with *equality of odds*, we want to study the asymptotic behavior of the estimators of the true positive and true negative rates across both groups. The reasoning is similar for the two rates, so we will only show the convergence of the true positive rate estimator, denoted in the following by $TP(g)$.

Theorem 2.7.2 *Set the following estimate of the true positive rate of a classifier g:*

$$R_n := \frac{\sum_{i=1}^n \mathbb{1}_{g(X_i)=1}\mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=0}\sum_{i=1}^n \mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=1}}{\sum_{i=1}^n \mathbb{1}_{g(X_i)=1}\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1}\sum_{i=1}^n \mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=0}}.$$

Then, the asymptotic distribution of this quantity is given by

$$\frac{\sqrt{n}}{\sigma}\left(R_n - TP(g)\right) \xrightarrow{d} N(0,1), \text{ as } n \to \infty, \tag{2.7.3}$$

where $\sigma = \sqrt{\nabla\varphi^T\left(\mathbb{E}Z_1\right)\Sigma_4\nabla\varphi\left(\mathbb{E}Z_1\right)}$ *and*

$$\nabla\varphi^T\left(\mathbb{E}Z_1\right) = \left(\frac{r_1}{p_1 r_0}, -\frac{p_0 r_1}{p_1^2 r_0}, -\frac{p_0 r_1}{p_1 r_0^2}, \frac{p_0}{p_1 r_0}\right)$$

$$\Sigma_4 = \begin{pmatrix} p_0(1-p_0) & & & \\ -p_0 r_1 & p_1(1-p_1) & & \\ p_0(1-r_0) & -p_1 r_0 & r_0(1-r_0) & \\ p_0 r_1 & p_1(1-r_1) & -r_0 r_1 & r_1(1-r_1) \end{pmatrix},$$

where we have denoted $p_s = \mathbb{P}(g(X_1) = 1, Y_1 = 1, S_1 = s)$, *and* $r_s = \mathbb{P}(Y_1 = 1, S_1 = s)$, *for* $s = 0, 1$.

Proof of Theorem 2.7.2 The proof follows the same guidelines of previous proof. We set here

$$Z_i = \begin{pmatrix} \mathbb{1}_{g(X_i)=1}\mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=0} \\ \mathbb{1}_{g(X_i)=1}\mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=1} \\ \mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=0} \\ \mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=1} \end{pmatrix},$$

where $\mathbb{1}_{g(X_i)=1}\mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=s} \sim B(\mathbb{P}(g(X_i)=1, Y_i=1, S_i=s))$ and $\mathbb{1}_{Y_i=1}\mathbb{1}_{S_i=s} \sim B(\mathbb{P}(Y_i=1, S_i=s))$, $s=0,1,$. From the Central Limit Theorem, we have that

$$\sqrt{n}\left(\bar{Z}_n - \mathbb{E}Z_1\right) \xrightarrow{d} N_4\left(\mathbf{0}, \Sigma_4\right), \ as \ n \to \infty.$$

with

$$\Sigma_4 = \begin{pmatrix} p_0(1-p_0) & & & \\ -p_0r_1 & p_1(1-p_1) & & \\ p_0(1-r_0) & -p_1r_0 & r_0(1-r_0) & \\ p_0r_1 & p_1(1-r_1) & -r_0r_1 & r_1(1-r_1) \end{pmatrix}. \tag{2.7.4}$$

Now consider the function

$$\begin{array}{rccc} \varphi: & \mathbb{R}^4 & \longrightarrow & \mathbb{R} \\ & (x_1,x_2,x_3,x_4) & \longmapsto & \dfrac{x_1x_4}{x_2x_3} \end{array}$$

Applying the Delta-Method for the function φ, we conclude that

$$\sqrt{n}\left(\varphi(\bar{Z}_n) - \varphi(\mathbb{E}Z_1)\right) \xrightarrow{d} \nabla\varphi^T\left(\mathbb{E}Z_1\right) N_4\left(\mathbf{0}, \Sigma_4\right), \ as \ n \to \infty,$$

where $\varphi(\bar{Z}_n) = R_n$, and $\varphi(\mathbb{E}Z_1) = TP(g)$. $\square$

2.7.3 Bootstraping vs. Direct Calculation of IC interval

The estimation of the Disparate Impact is unstable. In this paper we promote the use of the theoretical confidence interval based on the well known Delta method to control its variability. Contrary to Morris and Lobsenz, it does not rely on Gaussian approximation. We compare the stability of this confidence interval to bootstrap simulations, see for instance in Efron and Tibshirani [1994] for more details on bootstrap methods.

For this we build 1000 bootstrap replicates and estimate the disparate impact. Figure 2.9 presents the simulations. We can see that the bootstrap simulations remain in the confidence interval. Moreover, if we build a confidence interval for the bootstrap estimator, the confidence intervals are the same. We obtain by the theoretical confidence interval $[0.349, 0.384]$ while the bootstrap's confidence interval is $[0.349, 0.385]$. Hence the theoretical confidence is a reliable measure of fairness for the data set and should be preferred due to its small computation time compared to the 1000 bootstrap replication.

Note that in this paper, for sake of clarity, we have chosen to focus only on the disparate impact criterion. Yet all other fairness criteria should be given with the calculation of a confidence interval. For instance in del Barrio et al. [2019b] we propose confidence intervals for Wasserstein distance which is used in many methods in fair learning.

2.7.4 Application to other real datasets

To illustrate these tests we have also considered another two well-known and real data sets.

1. ***German Credit* data.** This data set is often claimed to exhibit some origin discrimination in the success of being given a credit by the German bank. Hence we compute the disparate impact w.r.t Origin. We obtain

   $$DI = 0.77 \in [0.68, 0.87].$$

   Hence here confidence intervals play an important role. Actually the disparate impact is not statistically significantly lower than 0.8, which entails that the discrimination of the decision rule of the German bank can not be shown, which promotes the use of a proper confidence interval.

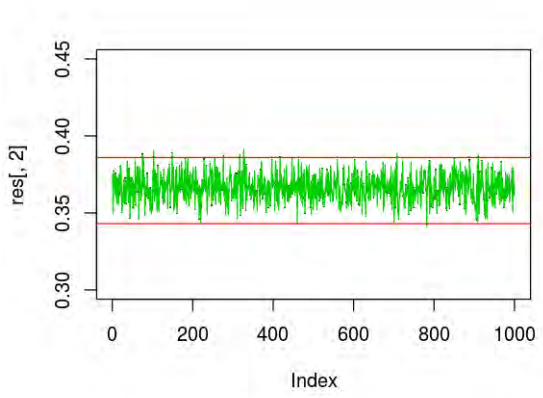


Figure 2.9 – Comparison with bootstrap computations

2. ***COMPAS Recidivism data***. A third data set is composed by the data of the controversial COMPAS score detailed in Dieterich et al. [2016]. The data is composed of 7214 offenders with personal variables observed over two years. A score predicts their level of dangerosity which determines whether they can be released while a variable points out if there has been recidivism. Hence Recidivism of offenders is predicted using a score and confronted to possible racial discrimination which corresponds to the protected attribute. The protected variable separates the population into caucasian and non caucasian. To evaluate the level of discrimination we first compute the disparate impact with respect to the true variable and the COMPAS score seen as a predictor.

$$DI = 0.76 \in [.72, .81]; \quad DI(\text{COMPAS}) = 0.71 \in [0.68; 0.74].$$

In both cases, the data are biased but the level of discrimination is low. Yet as mentioned in al the studies on this data set, the level of errors of prediction is significantly different according to the ethnic origin of the defender. Actually the conditional accuracy scores and their corresponding confidence intervals show clearly the unbalance treatment received by both populations.

$$TPR = 0.6 \in [0.54, 0.65]$$

$$TNR = 3.38 \in [2.46, 4.3]$$

This unbalanced treatment is clearly assessed with the confidence interval.

Chapter 3

Review of Mathematical Frameworks for Fairness in Machine Learning

The content of this chapter is available online in del Barrio et al. [2020] and currently submitted for publication.

Contents

A review of the main fairness definitions and fair learning methodologies proposed in the literature over the last years is presented from a mathematical point of view. Following our independence-based approach, we consider how to build fair algorithms and the consequences on the degradation of their performance compared to the possibly unfair case. This corresponds to the price for fairness given by the criteria *statistical parity* or *equality of odds*. Novel results giving the expressions of the optimal fair classifier and the optimal fair predictor (under a linear regression gaussian model) in the sense of *equality of odds* are presented.

3.1 Introduction

With both the introduction of new ways of storing, sharing and streaming data and the drastic development of the capacity of computers to handle large computations, the conception of models have changed. Mathematical models were first designed following prior ideas or conjectures from physical or biological models, then tested by designing experiments to test the validity of the ideas of their inventors. The model holds until new observations enable to reject its assumptions. The so-called Big Data's area introduced a new paradigm. The observed data convey enough information to understand the complexity of real life and the more the data, the better the description of the reality. Hence building models optimised to fit the data has become an efficient way to obtain generalizable models able to describe and forecast the real world.

In this framework, the principle of supervised machine learning is to build a decision rule from a set of labeled examples called the learning sample, that fits the data. This rule becomes a model or a decision algorithm that will be used for all the population. Mathematical guarantees can be provided in certain cases to control the generalization error of the algorithm which corresponds to the approximation done by building the model based on the observations and not knowing the true model that actually generated the data set. More precisely, the data are assumed to follow an unknown distribution while only its empirical distribution is at hand. So bounds are given to measure the error made by fitting a model on such observations and still using the model for new data. Yet the underlying assumption is that the observations follow all the same distribution which can be correctly estimated by the learning sample. Potential existing bias in the learning sample will be implicitly learnt and incorporated in the prediction. The danger of an uncontrolled prediction is greater when the algorithm lacks interpretability hence providing predictions that seem to be drawn from a yet accurate black-box but without any control or understanding on the reasons why they were chosen.

More precisely, in a supervised setting, the aim of a machine learning algorithm is to learn the relationships between characteristic variables X and a target variable Y in order to forecast new observations. Set the learning sample as $(Y_1, X_1), \dots, (Y_n, X_n)$ i.i.d observations drawn from an unknown distribution $\mathbb{P}$. Set the empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$. The quality of the prediction will be measured using a loss function defined as $\ell : (Y, \hat{Y}) \mapsto \ell(Y, \hat{Y}) \in \mathbb{R}^+$ to quantify the error made while predicting $\hat{Y}$ when Y is observed. Then for a given chosen class of algorithms $\mathcal{F}$, consider $\widehat{f}_n$ the best model that can be estimated by minimizing over $\mathcal{F}$, the loss function (and possibly a penalty to prevent overfitting for example), namely

$$\widehat{f}_n \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) + \lambda \text{penalty}(f) \right\}, \tag{3.1.1}$$

where λ balances the contribution of both terms to get a trade-off between the bias and the efficiency of the algorithm. The oracle rule is the best (yet unknown) rule that could be constructed if the true distribution were known

$$f^\star \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}\{\ell(Y, f(X)) + \lambda \text{penalty}(f)\}.$$

The predictions are given by $\widehat{Y} = \widehat{f}_n(X)$. Results from machine learning theory ensures that for proper choices of set of rules $\mathcal{F}$, the prediction's error behaves close to the oracle in the sense that, from a mathematical point of view, the excess risk

$$\mathbb{E}_{\mathbb{P}}\{\ell(Y, \widehat{f}_n(X))\} - \mathbb{E}_{\mathbb{P}}\{\ell(Y, f^\star(X))\}$$

is small. So mathematical guarantees warrant that the optimal forecast model reproduces the uses learnt from the learning set for new observations.

3.2 A definition of fairness in machine learning as independence criterion

3.2.1 Definition of full fairness

There is no doubt that machine learning is a powerful tool that is improving human life and has shown great promise in the developping of very different technological applications, including powering self-driving cars, accurately recognizing cancer in radiographs, or predicting our interests based upon past behavior, to name just a few. Yet with its benefits, machine learning also involves delicate issues such as the presence of bias in the model classifications and predictions. Hence, with this generalization of predictive algorithms in a wide variety of fields, algorithmic fairness is gaining more and more attention not only in the scientific research and Ethics communities (see for e.g. Besse et al. [2018a]), but also among the general population, who is experiencing a great impact on its daily life and activity. Thanks to this, there has been a push for the emergence of different approaches for assessing the presence of bias in machine learning algorithms over the last years. Similarly, various classifications have been proposed to understand the different sources of data bias. We refer to Mehrabi et al. [2019] for a recent review.

Consider the probability space $(\Omega, \mathcal{B}, \mathbb{P})$, with $\mathcal{B}$ the Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geq 1$. We will assume in the following that the bias is modeled by the random variable $S \in \mathcal{S}$ that represents an information about the observations $X \in \mathcal{X} \subset \mathbb{R}^d$, that should not be included in the model for the prediction of the target $Y \in \mathbb{R}^d,\ d \geq 1$. In the fair learning literature, the variable S is referred to as the *protected* or *sensitive attribute*. We assume moreover that this variable is observed. Most fairness theory has been developed particularly in the case when $\mathcal{S} = \{0, 1\}$ and S is a sensitive binary variable. In other words, the population is supposed to be possibly divided into two categories, taking the value $S = 0$ for the *minority* (assumed to be the unfavored class), and $S = 1$ for the *default* (and usually favored class). Hence, we also study more deeply this case and it will be conveniently indicated in the rest of the chapter, but in principle we consider general $\mathcal{S}$. From a mathematical point of view, we follow the recent paper Serrurier et al. [2019] that proposed the two following models that aim at understanding how this bias could be introduced in the algorithms:

1. The first model corresponds to the case where the data are subject to a bias nuisance variable which, in principle, is assumed not to be involved in the learning task, and whose influence in the prediction should be removed. We refer here to the well-known example of the dog vs. wolf in Ribeiro et al. [2016], where the input data were images highly biased by the presence of background snow in the pictures of wolves, and the absence of it in those of dogs. As shown in Figure 3.1a, this situation appears when the attributes X are a biased version of unobserved fair attributes $X^\star$ and the target variable Y depends only on $X^\star$. In this framework, learning from X induces biases while fairness requires:

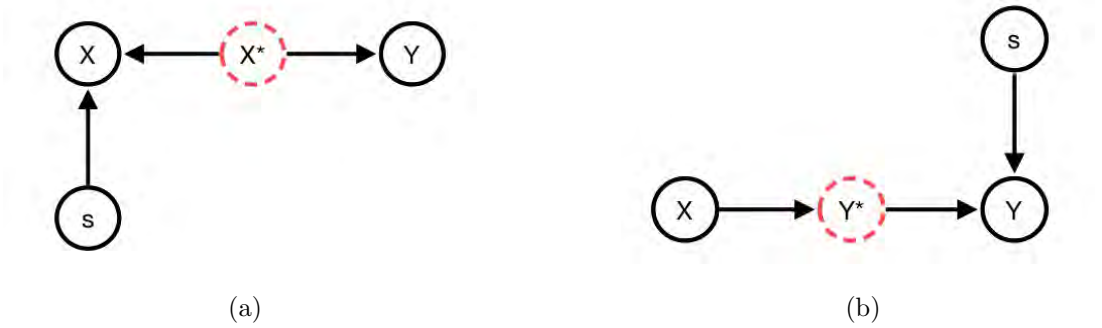
$$X^\star \perp\!\!\!\perp S \mid Y \quad \text{and} \quad Y \perp\!\!\!\perp S \mid X^\star.$$

   Note that neither X nor Y is independent of the protected S.

2. The second model corresponds to the situation when a biased decision is observed as a result of a fair score $Y^\star$ which has been biased by the uses giving rise to the target Y. Thus, a fair model in this case will change the prediction in order to make them independent of the protected variable. This is represented in Figure 3.1b and, formally, it is required that

$$X \perp\!\!\!\perp S \mid Y \quad \text{and} \quad Y^\star \perp\!\!\!\perp S \mid Y,$$

where $Y^\star$ is not observed. Note that previous conditions do not imply the independence between Y and S (even conditionally to X).



(a) (b)

Figure 3.1 – Two models for understanding the introduction of bias in the model

In the statistical literature, an algorithm $\hat{f}_n$ is called fair or unbiased when its outcome does not depend on the sensitive variable. The notion of *perfect fairness* requires that the protected variable S does not play any role in the forecast $\widehat{Y} = f(X, S)$ of the target Y. In other words, we will be looking at the independence between the protected variable S and the outcome $\widehat{Y}$, both considering given or not the true value of the target Y. These two notions of fairness are known in the literature as:

- *Statistical parity* (S.P.) deals with the independence between the outcome of the algorithm and the sensitive attribute
$$\hat{Y} \perp\!\!\!\perp S \tag{3.2.1}$$
- *Equality of odds* (E.O.) considers the independence between the protected attribute and the outcome conditionally given the true value of the target
$$\hat{Y} \perp\!\!\!\perp S \,|\, Y \tag{3.2.2}$$

Hence, a perfect fair model should be chosen within a class ensuring one of these restrictions (3.2.1)-(3.2.2). Observe that the choice of the notion of fairness is convenient regarding the assumed model for the introduction of the bias in the algorithm: while *statistical parity* is suitable for model 3.1a, *equality of odds* is for model 3.1b, and especially well-suited for scenarios where ground truth is available for historical decisions used during the training phase.

In this work, we tackle only these two main notions of fairness developed among the machine learning community. There are other definitions such as *avoiding disparate treatment* or *predictive parity*, defined respectively as $\hat{Y}|X \perp\!\!\!\perp S$ or $Y \perp\!\!\!\perp S|\hat{Y}$. A decision making system suffers from *disparate treatment* if it provides different outcomes for different groups of people with the same (or similar) values of non-sensitive features but different values of sensitive features [Barocas and Selbst, 2016]. In other words, (partly) basing the decision outcomes on the sensitive feature value amounts to disparate treatment. Technically, the *disparate treatment* doctrine tries to counter explicit as well as intentional discrimination [Barocas and Selbst, 2016]. It follows from the specification of *disparate treatment* that a decision maker with an intent to discriminate could try to disadvantage a group with a certain sensitive feature value (e.g., a specific race group) not by explicitly using the sensitive feature itself, but by intentionally basing decisions on a correlated feature (e.g., the non-sensitive feature location might be correlated with the sensitive feature race). This practice is often referred to as redlining in the US anti-discrimination law and also qualifies as disparate treatment [Gano, 2017]. However, such hidden intentional disparate

treatment maybe be hard to detect, and some authors argue that *statistical parity* might be a more suitable framework for detecting such covert discrimination [Siegel, 2014], while others focus only on explicit disparate treatment [Zafar et al., 2019]. For further details, we refer to the comprehensive study of fairness in machine learning given in Barocas et al. [2019].

The description of the metrics given above applies in a general context, yet all four fairness measures were originally proposed within the binary classification framework. Hence the literature cites and equivalent denominations will be presented in the following subsection specifically for this context.

3.2.2 The special case of classification

Fairness has been widely studied in the binary classification setting. Here the problem consists in forecasting a binary variable $Y \in \{0, 1\}$, using observed covariates $X \in \mathbb{R}^d,\ d \geq 1$. We introduce also a notion of positive prediction: $Y = 1$ represents a *success* while $Y = 0$ is a *failure*. We refer to Bousquet et al. [2004] for a complete description of classification problems in statistical learning. In this framework, the two main algorithmic fairness metrics are specified as follows.

- *Statistical parity.* Despite the early uses of this notion through the so-called $4/5^{th}$-rule for fair classification purposes by the State of California Fair Employment Practice Commission (FEPC) in 1971[1], it was first formally introduced as *statistical parity* in Dwork et al. [2012] in the particular case when S is also binary. Since then it has received several other denominations in the fair learning literature. For instance, it has been equivalently named in the same introductory work as *demographic parity* or *group fairness*, and also in others *equal acceptance rate* [Zliobaite, 2015] or *benchmarking* [Simoiu et al., 2017]. Formally, if $S \in \{0, 1\}$ this definition of fairness is satisfied when both subgroups are equally probable to have a successful outcome

$$\mathbb{P}(\hat{Y} = 1 \mid S = 0) = \mathbb{P}(\hat{Y} = 1 \mid S = 1), \tag{3.2.3}$$

  which can be extended to $\mathbb{P}(\hat{Y} = 1 \mid S) = \mathbb{P}(\hat{Y} = 1)$ for general S, continuous or discrete. A related and more rigid measure is called *avoiding disparate treatment* in Zafar et al. [2017a] if the probability that the classifier outputs a specific value of the forecast given a feature vector does not change after observing the sensitive feature, namely $\mathbb{P}(\hat{Y} = 1 \mid X, S) = \mathbb{P}(\hat{Y} = 1 \mid X)$.

- *Equality of odds* (or *equalized odds*) looks for the independence between the error of the algorithm and the protected variable. Hence, in practice, when S is also binary it compares the error rates of the algorithmic decisions between the different groups of the population, and considers that a classifier is fair when both classes have equal False and True Positive Rates

$$\mathbb{P}(\hat{Y} = 1 \mid Y = i, S = 0) = \mathbb{P}(\hat{Y} = 1 \mid Y = i, S = 1),\ \text{for } i = 0, 1. \tag{3.2.4}$$

  For general S, we note that this condition is equivalent to

$$\mathbb{P}(\hat{Y} = 1 \mid Y = i, S) = \mathbb{P}(\hat{Y} = 1 \mid Y = i),\ \text{for } i = 0, 1. \tag{3.2.5}$$

  This second point of view was introduced in Hardt et al. [2016] and has been originally proposed for recidivism of defendants in Flores et al. [2016]. Over the last few years it has been given several names, including *error rate balance* in Chouldechova [2017] or *conditional procedure accuracy equality* in Berk et al. [2018].

[1] https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

Many other metrics have received significant recent attention in the classification literature. In this setting, the already cited above *disparate treatment*, also referred to as *direct discrimination* [Pedreshi et al., 2008], looks at the equality for all $x \in \mathcal{X}$

$$\mathbb{P}(\hat{Y} = 1 \mid X = x, S = 0) = \mathbb{P}(\hat{Y} = 1 \mid X = x, S = 1) \tag{3.2.6}$$

Furthermore, we note that *equality of opportunity* (Hardt et al. [2016] or Kusner et al. [2017]) and *avoiding disparate mistreatment* [Zafar et al., 2017a] are two metrics related to the previous *equalized odds*, yet weaker. The first one requires only the equality of true positive rates, that is when $i = 1$ in (3.2.4), while the second looks at the equality of misclassification error rates across the groups:

$$\mathbb{P}(\hat{Y} \neq Y \mid S = 0) = \mathbb{P}(\hat{Y} \neq Y \mid S = 1). \tag{3.2.7}$$

Thus, *equality of odds* implies both the lack of *disparate mistreatment* and *equality of opportunity*, but not viceversa. Finally, we mention also here *predictive parity* which was introduced in Chouldechova [2017]. It requires the equality of positive predictive values across both groups. Therefore, mathematically it is satisfied when

$$\mathbb{P}(Y = 1 \mid \hat{Y} = 1, S = 0) = \mathbb{P}(Y = 1 \mid \hat{Y} = 1, S = 1). \tag{3.2.8}$$

The fairness metrics defined above are evaluated only for binary predictions and outcomes. By contrast, we can find also in the literature a set of metrics involving explicit generation of a continuous-valued score denoted here by $R \in [0, 1]$. Although scores could be used directly, they can alternatively serve as the input to a thresholding function that outputs a binary prediction.

Among this set, we highlight the notion of *test-fairness*, which extends *predictive parity* (3.2.8) when the prediction is a score. An algorithm satisfies this kind of fairness (or it is said to be *calibrated*) if for all scores r, the individuals who have the same score have the same probability of belonging to the positive class, regardless of group membership. Formally, this is expressed as $\mathbb{P}(Y = 1 \mid R = r, S = 0) = \mathbb{P}(Y = 1 \mid R = r, S = 1)$, for all scores r. This criteria was introduced in Chouldechova [2017] and has also been termed as *matching conditional frequencies* by Hardt et al. [2016].

A related metric called *well-calibration* [Verma and Rubin, 2018] or *calibration within groups* [Kleinberg et al., 2016] imposes an additional and more stringent condition: a model is *well-calibrated* if individuals assigned score r must have probability exactly r of belonging to the positive class. If this condition is satisfied, then *test-fairness* will also hold automatically, though not viceversa. Indeed, we note that the scores of a calibrated predictor can be transformed into scores satisfying well-calibration.

Finally, *balance for positive/negative class* was introduced in Kleinberg et al. [2016] as a generalization of the notion of *equality of odds*. Mathematically, this balance is expressed through the equalities of expected values $\mathbb{E}(R \mid Y = i, S = 0) = \mathbb{E}(R \mid Y = i, S = 1),\ i \in \{0, 1\}$.

3.2.3 Relationships between fairness criteria

It is also important to note that the wide variety of the proposed criteria formalizing different notions of fairness (see reviews Berk et al. [2018] and Verma and Rubin [2018] for more details) has led sometimes to incompatible formulations. The conditions under which more than one metric can be simultaneously satisfied, and relatedly, the ways in which different metrics might be in tension have been studied in several works [Chouldechova, 2017, Kleinberg et al., 2016, Berk et al., 2018]. Indeed, in the following Propositions 3.2.1, 3.2.2, 3.2.3 we revisit three *impossibility theorems of fairness* stating the exclusivity, except in non-degenerate cases, of the three main criteria considered in fair learning.

We study first the combination of all three of these metrics and then explore conditions under which it may be possible to simultaneously satisfy two metrics. To begin with, it is interesting to note that from the definition of conditional probability, the respective probability distributions associated with each of these three fairness metrics can be expressed as follows:

$$\mathcal{L}(Y, \hat{Y} \mid S) = \mathcal{L}(Y \mid \hat{Y}, S) \times \mathcal{L}(\hat{Y} \mid S) \tag{3.2.9}$$

$$= \mathcal{L}(\hat{Y} \mid Y, S) \times \mathcal{L}(Y \mid S). \tag{3.2.10}$$

We observe that on the right-hand side of equality (3.2.9) the first factor refers to *predictive parity*, while the second one to *statistical parity*. Similarly, in the equality (3.2.10) the first term represents *equality of odds* while the second one the base rate, that is the distribution of the true target among each group.

While the three results for fairness incompatibilities are stated hereafter in a general learning setting and their proofs are gathered in the Appendix 3.6.1, in this section we present a discussion in the binary classification framework. Let us consider then the following notations for $s \in \{0, 1\}$,

- $TPR_s := \mathbb{P}(\hat{Y} = 1 \mid Y = 1, S = s)$ the group-specific true positive rates
- $FPR_s := \mathbb{P}(\hat{Y} = 1 \mid Y = 0, S = s)$ the group-specific false positive rates
- $PPV_s := \mathbb{P}(Y = 1 \mid \hat{Y} = 1, S = s)$ the group-specific positive predictive values

We consider first if a predictor can simultaneously satisfy *equalized odds* and *statistical parity*.

Proposition 3.2.1 (Statistical parity vs. Equality of odds) *If S and Y are not independent and $\hat{Y}$ and Y are not independent, then statistical parity and equality of odds cannot hold simultaneously.*

In the special case of binary classification the result can be sharpened as follows. Observe that we can write for $s \in \{0, 1\}$,

$$\mathbb{P}(\hat{Y} = 1 \mid S = s) = \mathbb{P}(Y = 1 \mid S = s)TPR_s + \mathbb{P}(Y = 0 \mid S = s)FPR_s \tag{3.2.11}$$

Then computing the difference between expression (3.2.11) for each class and assuming that *equalized odds* holds, namely

$$TPR_0 = TPR_1 = \mathbb{P}(\hat{Y} = 1 \mid Y = 1) \text{ and } FPR_0 = FPR_1 = \mathbb{P}(\hat{Y} = 1 \mid Y = 0),$$

we obtain

$$\begin{aligned}
&\mathbb{P}(\hat{Y} = 1 \mid S = 0) - \mathbb{P}(\hat{Y} = 1 \mid S = 1) \\
&= (\mathbb{P}(Y = 1 \mid S = 0) - \mathbb{P}(Y = 1 \mid S = 1))\mathbb{P}(\hat{Y} = 1 \mid Y = 1) \\
&+ (\mathbb{P}(Y = 0 \mid S = 0) - \mathbb{P}(Y = 0 \mid S = 1))\mathbb{P}(\hat{Y} = 1 \mid Y = 0) \\
&= (\mathbb{P}(Y = 1 \mid S = 0) - \mathbb{P}(Y = 1 \mid S = 1))(\mathbb{P}(\hat{Y} = 1 \mid Y = 1) - \mathbb{P}(\hat{Y} = 1 \mid Y = 0))
\end{aligned}$$

Statistical parity requires that left side is exactly zero. Hence, for the right side also being zero necessarily $\mathbb{P}(Y = 1 \mid S = 0) = \mathbb{P}(Y = 1 \mid S = 1)$ or $\mathbb{P}(\hat{Y} = 1 \mid Y = 1) = \mathbb{P}(\hat{Y} = 1 \mid Y = 0)$. However, it is usually assumed that base rates differs across the groups, that is, the ratio of people in the group who belong to the positive class ($Y = 1$) to the total number of people in that group. Thus, *statistical parity* and *equalized odds* are simultaneously achieved only if true and false positive rates are equal. While this is mathematically possible, such condition is not particularly useful since the goal is typically to develop a predictor in which the true positive rate is significantly higher than the false.

Proposition 3.2.2 (Statistical parity vs. Predictive parity) *If S and Y are not independent, then statistical parity and predictive parity cannot hold simultaneously.*

By contrast, in the binary classification setup the two fairness metrics are actually simultaneously feasible. Assume that *statistical parity* holds, that is, $\mathbb{P}(\hat{Y} = 1|S = 1) = \mathbb{P}(\hat{Y} = 1|S = 0) = \mathbb{P}(\hat{Y} = 1)$. Then, from equations (3.2.9)-(3.2.10) we can write the difference of positive predictive values

$$PPV_0 - PPV_1 = \frac{TPR_0\mathbb{P}(Y = 1|S = 0) - TPR_1\mathbb{P}(Y = 1|S = 1)}{\mathbb{P}(\hat{Y} = 1)} \tag{3.2.12}$$

Under *predictive parity* the left side of the above equation must be zero, which in turn requires that the ratio of the true positive rates of the two groups be the reciprocal of the ratio of the base rates, namely

$$\frac{TPR_0}{TPR_1} = \frac{\mathbb{P}(Y = 1|S = 1)}{\mathbb{P}(Y = 1|S = 0)} \tag{3.2.13}$$

Thus, while statistical and predictive parity can be simultaneously satisfied even with different base rates, the utility of such a predictor is limited when the ratio of the base rates differs significantly from 1, as this forces the true positive rate for one of the groups to be very low.

Proposition 3.2.3 (Predictive parity vs. Equality of odds) *If S and Y are not independent then predictive parity and equality of odds cannot hold simultaneously.*

We explore this incompatibility in more detail in the binary classification framework. If both conditions hold

$$TPR_0 = TPR_1, \quad FPR_0 = FPR_1, \quad \text{and} \quad PPV_0 = PPV_1, \tag{3.2.14}$$

so we can write

$$\mathbb{P}(\hat{Y} = 1 \mid S) = \sum_{i=0,1} \mathbb{P}(\hat{Y} = 1 \mid Y = i, S)\mathbb{P}(\hat{Y} = 1 \mid S) = TPR_0\mathbb{P}(Y = 1 \mid S) + FPR_0\mathbb{P}(Y = 0 \mid S).$$

This together with equations (3.2.9)-(3.2.10) implies

$$\begin{aligned}&\mathbb{P}(\hat{Y} = 1|Y = 1, S = 0)\mathbb{P}(Y = 1|S = 0)\\ &= \mathbb{P}(y = 1|\hat{Y} = 1, S = 0)\left[TPR_0\mathbb{P}(\hat{Y} = 1|S) + FPR_0\mathbb{P}(Y = 0|S = 0)\right],\end{aligned}$$

and using the notations above we obtain

$$TPR_0\mathbb{P}(Y = 1|S = 0) = PPV_0\left[TPR_0\mathbb{P}(\hat{Y} = 1|S) + FPR_0(1 - \mathbb{P}(Y = 1|S = 0))\right].$$

Finally, we obtain the following expressions for the group-specific base rate for $s = 0$

$$\mathbb{P}(Y = 1|S = 0) = \frac{PPV_0FPR_0}{PPV_0FPR_0 + (1 - PPV_0)TPR_0} \tag{3.2.15}$$

and reasoning likewise for $s = 1$

$$\mathbb{P}(Y = 1|S = 1) = \frac{PPV_1FPR_1}{PPV_1FPR_1 + (1 - PPV_1)TPR_1} \tag{3.2.16}$$

Hence, in the absence of perfect prediction, under assumption (3.2.14) base rates have to be equal for both *equalized odds* and *predictive parity* to simultaneously hold. When perfect prediction is achieved, equations (3.2.15) and (3.2.16) take on the indefinite form $0/0$ so therefore do not convey anything definitive about base rates in that scenario.

We also note that the less strict metric *equal opportunity* (recall it requires only equal TPR across groups) is compatible with *predictive parity*. This is evident from equations (3.2.15) and (3.2.16) when the condition $FPR_0 = FPR_1$ is removed, thereby allowing *equalized opportunity* and *predictive parity* to be simultaneously satisfied even with unequal base rates. However, achieving this condition with unequal base rates will require that the FPR differs across the groups. When the difference between the base rates is large, the variation between group-specific FPRs may have to be significant which may reduce suitability for some applications. Hence, while *equal opportunity* and *predictive parity* are compatible in the presence of unequal base rates, practitioners should consider the cost (in terms of FPR difference) before attempting to simultaneously achieve both. A similar analysis is possible when we considering parity in negative predictive value instead of positive predictive value, i.e. equal opportunity and parity in NPV are compatible, but only at the cost of variation between group-specific true negative rates (TNRs).

3.3 Price for fairness in machine learning

In this section, we consider how to build fair algorithms and the consequences on the degradation of their performance compared to the *possibly unfair* case. This corresponds to the price for fairness.

Recall that the performance of an algorithm is measured through its risk defined by

$$R(f) = \mathbb{E}(\ell(Y, f(X, S))).$$

Define some class or restriction of classes

$$\mathcal{F}_{SP} = \{f(X, S) \in \mathcal{F} \quad \text{s.t} \quad \hat{Y} \perp\!\!\!\perp S\} \tag{3.3.1}$$

$$\mathcal{F}_{EO} = \{f(X, S) \in \mathcal{F} \quad \text{s.t} \quad \hat{Y}|Y \perp\!\!\!\perp S\} \tag{3.3.2}$$

From a theoretical point of view, a fair model can be achieved by restricting the minimization (3.1.1) to such classes. The price for fairness is

$$\mathcal{E}_{\text{Fair}}(\mathcal{F}) := \inf_{f \in \mathcal{F}_{\text{Fair}}} R(f) - \inf_{f \in \mathcal{F}} R(f). \tag{3.3.3}$$

If $\mathcal{F}$ denotes the class of all measurable functions, then $\inf_{f \in \mathcal{F}} R(f)$ is known as the Bayes Risk. In the following, we will study the difference of the minimal risks in (3.3.3) under both fairness assumptions and in two different frameworks, regression and classification, through an optimal transport based approach.

Optimal transport (OT) is a foundational problem in optimization, that allows to compare probability distributions while taking into account geometric aspects. Its optimal objective value, the Wasserstein (a.k.a Monge-Kantorovich) distance, provides an important loss between distributions that has been used in many applications throughout machine learning and statistics, being one of the current trends among the research community [Hütter and Rigollet, 2019, Bigot, 2019, Ballu et al., 2020, Mérigot et al., 2019, Niles-Weed and Rigollet, 2019]. We refer to Villani [2009] for a detailed description on OT theory. For P and Q two probability measures

on $\mathbb{R}^d$, the squared Wasserstein distance between P and Q is defined as

$$\mathcal{W}_2^2(P,Q) := \min_{\pi \in \Pi(P,Q)} \int \|x-y\|^2 d\pi(x,y)$$

where $\Pi(P,Q)$ the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q.

3.3.1 Price for fairness as Statistical Parity

The notion of perfect fairness given by *statistical parity* criterion implies that the distribution of the predictor does not depend on the protected variable S.

3.3.1.1 Regression

In the regression problem, *statistical parity* condition is expressed through the equality of distributions $\mathcal{L}(f(X,S)|S) = \mathcal{L}(f(X,S))$. Then in this setting a standard definition of this statistical independence requires that $\mathbb{P}(f(X,s) \in A | S = s) = \mathbb{P}(f(X,S) \in A)$ for all $s \in \mathcal{S}$ and all measurable sets A. Since $f(X,S)$ is a real-valued random variable under Borel σ-algebra, it is fully characterized by its cumulative distribution function, and so it suffices to consider sets $A = [x, +\infty)$, for $x \in \mathbb{R}$.

This fairness assumption implies the weakest cases where $\mathbb{E}(f(X,S)|S) = \mathbb{E}(f(X,S))$ as presented in the works of Dwork et al. [2012] and Zemel et al. [2013], or equivalently when $\mathrm{Cov}(f(X,S),S) = 0$. Note that in the case where S is a discrete variable, the previous criteria have a simpler expression. In particular, in the binary setup when $S \in \{0,1\}$, we can write

$$\begin{aligned}\mathbb{E}_{X,S}(f(X,S)) &= \mathbb{E}_X[\mathbb{E}_S[f(X,S) \mid S]] \\ &= \mathbb{P}(S=0)\mathbb{E}_X(f(X,0) \mid S=0) + \mathbb{P}(S=1)\mathbb{E}_X(f(X,1) \mid S=1).\end{aligned}$$

Then we have that *statistical parity* holds if, and only if,

$$\mathbb{E}_X(f(X,S) \mid S=0) = \mathbb{E}_X(f(X,S) \mid S=1).$$

In the general regression setting, we will use the following notations : $X \in \mathcal{X}$, $S \in \mathcal{S}$, $Y \in \mathbb{R}^d$ When $\mathcal{F}$ is the set of all measurable functions from $\mathcal{X} \times \mathcal{S}$ to $\mathbb{R}^d$, the optimal risk (a.k.a. Bayesian risk), is defined as

$$R^\star := \mathcal{R}(\mathcal{F}) = \min_{f \in \mathcal{F}} \mathbb{E}\|Y - f(X,S)\|^2$$

is achieved for the Bayes estimator

$$\eta(X,S) := \mathbb{E}[Y|(X,S)].$$

Denote μ_S the conditional distribution of the Bayes estimator $\mathbb{E}(Y|X,S)$ given S and for a predictor g $\nu_S(g)$ the conditional distribution of $g(X,S)$ given S. In Le Gouic and Loubes [2020] the authors relate the excess risk with a minimization problem in the Wasserstein space proving the following lower bound for the price for fairness.

Theorem 3.3.1

$$\inf_{f \in \mathcal{F}_{\text{Fair}}} R(f) - \inf_f R(f) \geq \inf_{g \in \mathcal{F}} \mathbb{E}\mathcal{W}_2^2(\mu_S, \nu_S(g)). \tag{3.3.4}$$

Moreover, if $\mathcal{F} = \mathcal{F}_{SP}$ and μ_s has density w.r.t. Lebesgue measure for almost every s, then (3.3.4) *becomes an equality*

$$\mathcal{E}_{\text{Fair}}(\mathcal{F}) = \inf_{g \in \mathcal{F}} \mathbb{E}_S \mathcal{W}_2^2(\mu_S, \nu_S(g)). \tag{3.3.5}$$

Imposing fairness comes at a price that can be quantified which depends on the 2-Wasserstein distance between distributions of Bayes predictors.

Finding the minimum in (3.3.5) is related to the minimization of Wasserstein's variation which has been known as the problem of studying Wasserstein's barycenter. Actually, for *statistical parity* constraint

$$\inf_{g\in\mathcal{F}} \mathbb{E}_S \mathcal{W}_2^2(\mu_S, \nu_S(g)) = \inf_{\nu(g)} \mathbb{E}_S \mathcal{W}_2^2(\mu_S, \nu(g))$$

which amounts to minimize

$$\nu \mapsto \mathbb{E}_S \mathcal{W}_2^2(\mu_S, \nu)$$

This problem has been studied in Agueh and Carlier [2011], Le Gouic and Loubes [2017] or del Barrio and Loubes [2019]. The distributions μ_S are random distributions and define $\mathbb{P}_{\mathcal{S}}$ their distribution on the set of distributions. Hence The minimum is reached for μ_B the Wasserstein barycenter of $\mathbb{P}_{\mathcal{S}}$. Note that if S is discrete, in particular for the two class version $S \in \{0, 1\}$, note $\pi_s = P(S = s)$, the distribution $\mathbb{P}_{\mathcal{S}}$ can be written as $\mathbb{P}_{\mathcal{S}} = \pi_1 \delta_{\mu_1} + (1 - \pi_1)\delta_{\mu_0}$. Hence its barycenter is a measure that minimizes the functional

$$\nu \mapsto \pi_0 \mathcal{W}_2^2(\mu_0, \nu) + (1 - \pi_0)\mathcal{W}_2^2(\mu_1, \nu).$$

Existence and uniqueness are ensured as soon as the μ_S have density with respect to Lebesgue measure.

3.3.1.2 Classification

We consider the problem of quantifying the price for imposing *statistical parity* when the goal is predicting a label. In the following and without loss of generality, we assume that Y is a binary variable with values in $\{0, 1\}$. If S is also binary, then Statistical Parity is often quantified in the fair learning literature using the so-called Disparate Impact (DI)

$$DI(g, X, S) = \frac{\mathbb{P}(g(X, S) = 1 \mid S = 0)}{\mathbb{P}(g(X, S) = 1 \mid S = 1)}. \tag{3.3.6}$$

This measures the risk of discrimination when using the decision rule encoded in g on data following the same distribution as in the test set. Hence, in Gordaliza et al. [2019] a classifier g is said not to have a Disparate Impact at level $\tau \in (0, 1]$ when $DI(g, X, S) > \tau$. Perfect fairness is thus equivalent to the assumption that the disparate impact is exactly $DI(g, X, S) = 1$. Note that the notion of DI defined Eq. (3.3.6) was first introduced as the $4/5^{th}$-rule by the State of California Fair Employment Practice Commission (FEPC) in 1971. Since then, the threshold $\tau_0 = 0.8$ was chosen in different trials as a legal score to judge whether the discriminations committed by an algorithm are acceptable or not (see e.g. Feldman et al. [2015] Zafar et al. [2017a], or Mercat-Bruns [2016]).

While in the classification problem the notion of *statistical parity* can be easily extended for general $S \in \mathcal{S}$, continuous or discrete, through the equality $\mathbb{P}(g(X, S) = 1) = \mathbb{P}(g(X, S) = 1 \mid S)$, the index Disparate Impact has not been used in the literature for quantifying fairness in the general framework. Hence, we only consider the classification problem. Still, if S is a multiclass sensitive variable, we observe that a fair classifier should satisfy for all $s \in \mathcal{S}$,

$$\mathbb{P}(g(X, S) = 1) = \mathbb{P}(g(X, S) = 1 \mid S = s). \tag{3.3.7}$$

Hence, Disparate Impact could be extended to

$$DI(g, X, S) = \frac{\min_{s\in\mathcal{S}} \mathbb{P}(g(X, S) = 1 \mid S = s)}{\mathbb{P}(g(X, S) = 1 \mid S = 1)}. \tag{3.3.8}$$

Tackling the issue of computing a bound in (3.3.3) is a difficult task and has been studied by several authors. In this specific framework, finding a lower bound for the loss of accuracy induced by the full statistical parity constraint has not been solved. This is mainly due to the fact that the classification setting does not specify a model to constrain the relationships between the labels Y and the observations X, enabling a too large choice of models, contrary to the regression case.

Yet in different frameworks, some results can be proved. On the one hand, in Jiang et al. [2019] a notion of fairness is considered which correspond to controling the number of class changes when switching labels, which amounts to study the difference between classification errors for plug in rules corresponding to all possible thresholds τ of Bayes score called the model belief, $\eta_S(X) = P(Y = 1|X, S) \geq \tau$. Authors achieve a bound using the W_1 distance and prove that the minimum loss is achieved for the 1-Wasserstein barycenter.

In the following we recall results obtained in Gordaliza et al. [2019] which study the price for fairness in statistical parity in the framework where we want to ensure that all classifiers trained by a transformation of the data will be fair with respect to the statistical parity definition.

For this consider the Balanced Error Rate

$$BER(g, X, S) = \frac{\mathbb{P}\left(g(X, S) = 0 \mid S = 1\right) + \mathbb{P}\left(g(X, S) = 1 \mid S = 0\right)}{2}$$

corresponding to the problem of estimating the sensitive label from the prediction in the most difficult case where the class are well balanced between each group labeled by the variable S. In this setting, unpredictability of the label warrants the fairness of the procedure. Actually, given $\varepsilon > 0, S$ is not $\varepsilon-$predictable from X if $BER(g, X, S) > \varepsilon$, for all $g \in \mathcal{G}$

$$DI(g, X, S) := \frac{a(g)}{b(g)}.$$

We consider classifiers g such that $a(g) > 0$ and $b(g) > 0$.

Theorem 3.3.2 (Link between Disparate Impact and Predictability) *Given random variables $X : \Omega \to \mathbb{R}^d$, $S : \Omega \to \{0, 1\}$, the classifier $g \in \mathcal{G}$ has Disparate Impact at level $\tau \in [0, 1]$, with respect to (X, S), if, and only if, S is $\left(\frac{1}{2} - \frac{a(g)}{2}(\frac{1}{\tau} - 1)\right)-$predictable from X.*

Then, we can see that the notion of predictability and the distance in Total Variation between the conditional distributions of $X \mid S$ are connected through the following theorem

Theorem 3.3.3 (Total Variation distance) *Given the variables $X : \Omega \to \mathbb{R}^d$, $d \geqslant 1$, and $S : \Omega \to \{0, 1\}$,*

$$\min_{g \in \mathcal{F}} BER(g, X, S) = \frac{1}{2}\left(1 - d_{TV}\left(\mathcal{L}\left(X|S = 0\right), \mathcal{L}\left(X|S = 1\right)\right)\right),$$

where $g : \mathbb{R}^d \to \{0, 1\}$ varies in the family of binary classifiers $\mathcal{G}$.

S is not $\varepsilon-$predictable from X if

$$d_{TV}\left(\mathcal{L}\left(X|S = 0\right), \mathcal{L}\left(X|S = 1\right)\right) < 1 - 2\varepsilon$$

where d_{TV} is the Total Variation distance.

Hence fairness for all classifier f is equivalent to the fact that

$$\min_{g \in \mathcal{F}} BER(g, X, S) = \frac{1}{2}$$

which is equivalent to

$$d_{TV}(\mu_0, \mu_1) = 0,$$

where we have set $\mu_S = \mathcal{L}(X|S)$ for $S \in \{0, 1\}$. Hence, perfect fairness for all classifiers in classification is equivalent to the fact that the distance between conditional distributions of the characteristics of individuals for the class defined by the different values of S is null.

Consider transformations that map the conditional distributions to a joint distribution. Consider $X \in \mathbb{R}^d$ and $S \in \{0, 1\}$. Let $T_S : \mathbb{R}^d \to \mathbb{R}^d,\ d \geqslant 1$ be a random transformation of X such that $\mathcal{L}(T_0(X) \mid S = 0) = \mathcal{L}(T_1(X) \mid S = 1)$, and consider the transformed version $\tilde{X} = T_S(X)$. This transformation defines a way to *repair* the data in order to achieve fairness for all possible classifiers applied to these repaired data $\tilde{X} = T_S(X)$. This maps transforms the distributions μ_S into their image by T_S, namely for all $S \in \{0, 1\}$, $\mu_{S\sharp}T_S := \mu_S \circ T_S^{-1}$. Note that the choice of the transformation is equivalent to the choice of the target distribution $\nu_S = \mu_{S\sharp}T_S$. Fairness is then achieved when the distance in Total Variations is equal to zero, which amounts to say that T_0 and T_1 maps the conditional distributions towards thew same distributions, hence $\nu_0 = \nu_1$. In this framework the price of fairness can be quantified as follows. For a given deformation T_S, set

$$\mathcal{E}(T_S) := \inf_{g \in \mathcal{G}} P(g(\tilde{X}) \neq Y) - R_B(X, S).$$

The following theorem provides an upper bound for this price for fairness.

Theorem 3.3.4 *(Gordaliza et al. [2019]) For each $s \in \{0, 1\}$, assume that the function $\eta_s(x) = \mathbb{P}(Y = 1 \mid X = x, S = s)$ is Lipschitz with constant $K_s > 0$. Then, if $K = \max\{K_0, K_1\}$,*

$$\mathcal{E}(T_S) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_{s\sharp}T_s) \right)^{\frac{1}{2}}.$$

Hence the minimal excess risk in this setting is achieved by minimizing previous quantity over possible transformations T_S. We thus obtain the following upper bound.

$$\begin{aligned}
\inf_{T_S} \mathcal{E}(T_S) &\leq 2\sqrt{2}K \inf_{T_S} \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_{s\sharp}T_s) \right)^{\frac{1}{2}} \\
&\leq 2\sqrt{2}K \inf_{\nu} \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \nu) \right)^{\frac{1}{2}} \\
&= \sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_B) \right)^{\frac{1}{2}}
\end{aligned}$$

where μ_B denotes the Wasserstein barycenter between μ_S with weight π_S for $S \in \{0, 1\}$.
Note that previous theorem can easily be extended to the case where S takes multiple discrete values $S \in \{1, \ldots, k\}$. In the case where S is continuous, the same result holds using the extension of Wasserstein barycenter in Le Gouic and Loubes [2017] and provided that conditional distributions μ_S are absolutely continuous with respect to Lebesgue measure.

3.3.2 Price for fairness as Equality of Odds

We study now the price for fairness meant as *equality of odds*, which looks at the independence between the protected attribute and the outcome conditionally given the true value of the target, that is, the error of the algorithm.

3.3.2.1 Regression

Consider the regression framework detailed in section 3.3.1.1 and let $(X_1, S_1, Y_1), \ldots, (X_n, S_n, Y_n)$ be a sample of i.i.d. random vectors observed from (X, S, Y). Denote by $\mathbb{X} \in \mathbb{R}^{n \times p}$ and $\mathbb{S} \in \mathbb{R}^{n \times 1}$ the matrices containing the observations of the non-sensitive and sensitive, respectively, features X and S. We will assume standard normal independent errors $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, 1)$. Then, we consider the linear normal model

$$Y = f_{\beta_0,\beta}(X, S) + \varepsilon, \tag{3.3.9}$$

where the errors are such that $\mathbb{E}(\varepsilon \mid (X, S)) = 0$, and the predictor

$$f_{\beta_0,\beta}(X, S) = \beta_0 S + \beta^T X, \ \beta_0 \in \mathbb{R}, \ \beta \in \mathbb{R}^{p \times 1} \tag{3.3.10}$$

is a linear combination of the sensitive and non-sensitive attributes. Then, the joint distribution of (X, S, Y) is $(p+2)-$dimensional normal and we denote the vectors of means and the covariance matrices as follows

$$(X, S, Y) \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_S \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_{XS} & \Sigma_{XY} \\ \Sigma_{XS}^T & \Sigma_S & \Sigma_{SY} \\ \Sigma_{XY}^T & \Sigma_{SY}^T & \Sigma_Y \end{bmatrix} \right)$$

We note that the *equality of odds* criterion requires the linear fair predictor being independent of S conditionally given Y, that is

$$f_{\beta_0,\beta}(X, S) \perp\!\!\!\perp S \mid Y,$$

which under the normal model is equivalent to the second order moment constraint

$$Cov(f(X, S), S \mid Y) = 0. \tag{3.3.11}$$

Hence, seeking for a fair linear predictor amounts to obtaining conditions on the coefficients β_0, β for (3.3.11) to hold. Since linear prediction can be seen as the most suitable framework for Gaussian processes, the relaxation of (3.3.11) could be justified as being the appropriate notion of fairness when we restrict ourselves to linear predictors. Furthermote, linear predictors, especially under kernel transformations, are used in a wide array of applications. They thus form a practically relevant family of predictors where one would like to achieve non-discrimination. Therefore, in this section, we focus on obtaining non-discriminating linear predictors.

Now if we denote by $C_{S,X,Y} \in \mathbb{R}^{p \times 1}$ the *vector of correction for fairness*

$$C_{S,X,Y} := \left(\frac{\Sigma_{XS}\Sigma_Y - \Sigma_{SY}\Sigma_{XY}}{\Sigma_S\Sigma_Y - \Sigma_{SY}^2} \right), \tag{3.3.12}$$

then the optimal fair *equality of odds* predictor under the normal model can be exactly computed as in the following result, whose proof is set out in the Appendix 3.6.2.

Proposition 3.3.5 *Under the normal model* (3.3.9)*, the optimal fair (equality of odds) linear predictor of the form* (3.3.10) *is given as the solution to the following optimization problem*

$$\left(\hat{\beta}_{0,fair}, \hat{\beta}_{fair}\right) := argmin_{(\beta_0,\beta) \in \mathcal{F}_{EO}} \mathbb{E}\left[(Y - f_{\beta_0,\beta}(X, S))^2 \right]$$
$$\mathcal{F}_{EO} = \{(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p \text{ such that } \beta^T \left(\Sigma_{XS}\Sigma_Y - \Sigma_{SY}\Sigma_{XY}\right) + \beta_0 \left(\Sigma_S\Sigma_Y - \Sigma_{SY}^2\right) = 0\}.$$

If moreover Y and S are not linearly dependent, it can be exactly computed as

$$\hat{\beta}_{0,fair} = \hat{\beta}_{fair}^T C_{S,X,Y}$$
$$\hat{\beta}_{fair} = \Sigma_Z^{-1}\Sigma_{ZY},$$

where

$$\Sigma_Z = \Sigma_X + \Sigma_S C_{S,X,Y} C_{S,X,Y}^T + C_{S,X,Y}\Sigma_{XS}^T + \Sigma_{XS} C_{S,X,Y}^T$$
$$\Sigma_{ZY} = \Sigma_{XY} + \Sigma_{SY} C_{S,X,Y}.$$

Note also that the case where Y and S are linearly dependent corresponds to a totally unfair scenario that is not worth studying. This result shows that, under the normal model, it is possible to quantify the excess of risk attributable to achieving a fair regression. Precisely, we can compute the loss when imposing the *equality of odds* condition $(\beta_0, \beta) \in \mathcal{F}_{EO}$ by comparing with the general loss associated to the minimizer

$$\left[\hat{\beta}_0, \hat{\beta}^T\right]^T := argmin_{(\beta_0,\beta)\in\mathbb{R}\times\mathbb{R}^p} \mathbb{E}\left[(Y - f_{\beta_0,\beta}(X,S))^2\right]. \tag{3.3.13}$$

We have performed some simulations to obtain estimations of the minimal excess risk in (3.3.3) when imposing *equality of odds* under this gaussian linear regression framework. Precisely, we have considered $S \sim \mathcal{N}(0, 10)$ and $X \in \mathbb{R}^2$, such that

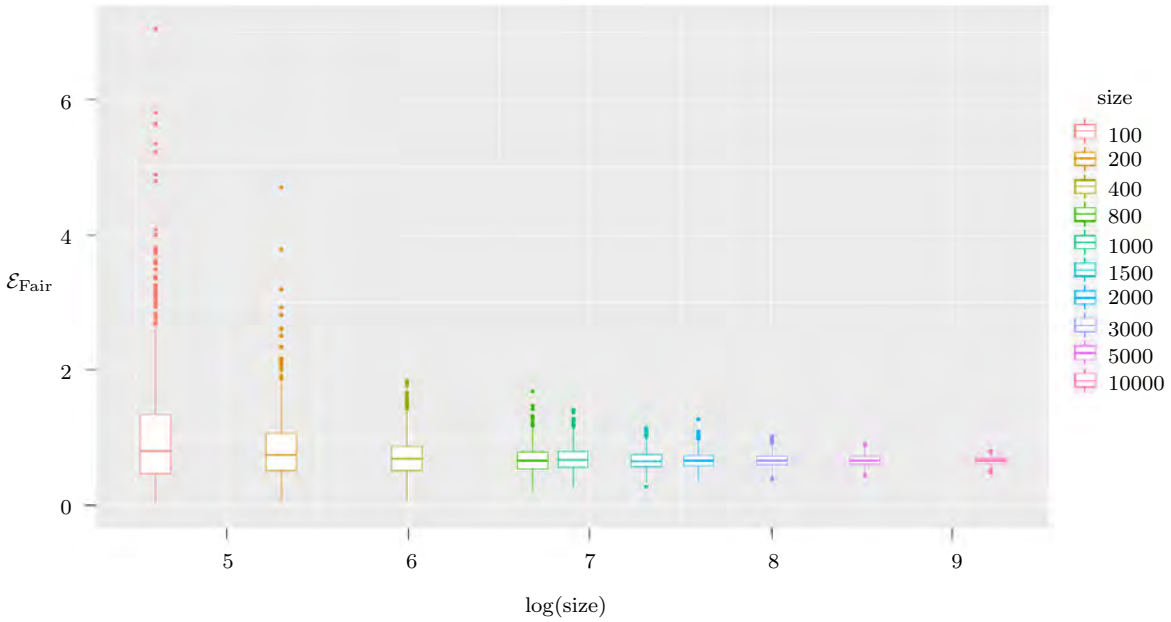
$$X \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\right).$$

The results of 1000 replications of the experiment are shown in Figure 3.2. There we present: (a) the average minimal excess risk; and its (b) standard deviation, as the sample size increases, taking particularly the values $(100, 200, 400, 800, 1000, 1500, 2000, 3000, 5000, 10000)$. We observe that the estimation seems to converge.

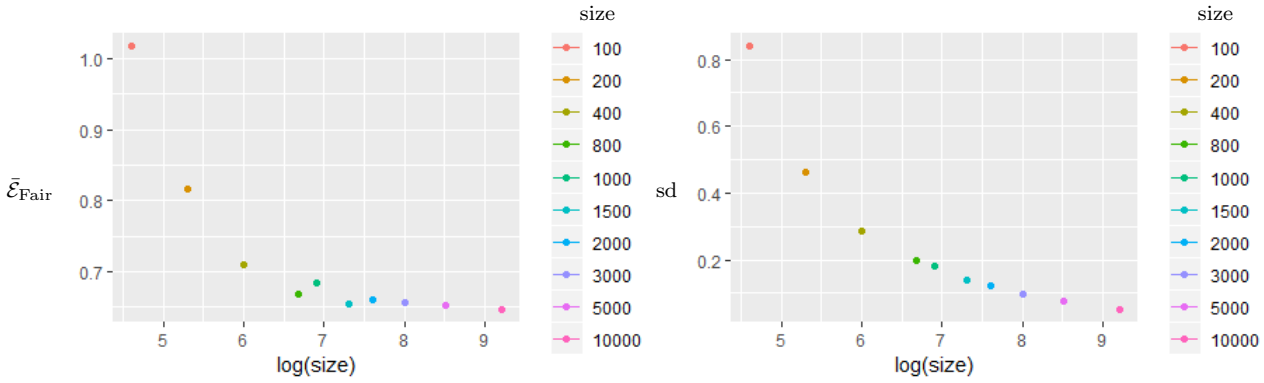
Moreover we observe that, while condition (3.3.11) is equivalent to *equality of odds* in the normal setting, it is generally a weaker constraint. However, the problem of achieving perfect fairness as *equalized odds* in a wider setup conveys computational challenges as discussed in Woodworth et al. [2017]. They showed that even in the restricted case of learning linear predictors, assuming a convex loss function, and demanding that only the sign of the predictor needs to be non-discriminatory, the problem of matching FPR and FNR requires exponential time to solve in the worst case. Motivated by this hardness result (see Theorem 3 in Woodworth et al. [2017]), they also proposed a relaxation of the criterion of *equalized odds* by a more tractable notion of non-discrimination based on second order moments. In particular, they proposed the notion of *equalized correlations*, which indeed is generally a weaker condition than (3.3.11), but when considering the squared loss and when (X, S, Y) are jointly Gaussian, it is in fact equivalent (and, subsequently, equivalent to *equality of odds*). They also point out that for many distributions and hypothesis classes, there may not exist a non-constant, deterministic, perfectly fair predictor. Hence, we have restricted ourselves here to the normal framework in which the computation of the optimal fair predictor is still feasible.

3.3.2.2 Classification

We consider again the classification setting where we wish to predict a binary output label $Y \in \{0, 1\}$ from the pair (X, S). In this section, we obtain the fair optimal classifier in the sense of *equality of odds* in the particular case where S is also binary. We assume moreover that both the marginals and the joint distribution of (S, Y) are non-degenerate, that is $\mathbb{P}(Y = 1) \in$



(a) Boxplot of the minimal excess risk computations



(b) average minimal excess risk

(c) standard deviation

Figure 3.2 – Minimal excess risk with $Cov(X_1, S) = 0.1,\ Cov(X_2, S) = 0.1$

$(0, 1)$, $\mathbb{P}(S = 1) \in (0, 1)$ and $\mathbb{P}(Y = 1, S = 1) \in (0, 1)$. There are some other works dealing with the computation of Bayes-optimal classifiers under different notions of fairness. In Menon and Williamson [2018] *statistical parity* and *equality of oportunity* are the considered constraints. Our approach here extends the proposed in Chzhen et al. [2019], where fairness is defined by the weaker notion of *equality of opportunity* that requires just the equality of true posisitive rates across both groups.

An optimal fair classifier is formally defined here as the solution to the risk minimination problem over the class $\mathcal{F}_{EO}$ of binary classfiers satisfying the *equality of odds* conditions, that is

$$g^* \in argmin_{g \in \mathcal{F}_{EO}} \mathcal{R}(g), \text{ where}$$
$$\mathcal{F}_{EO} := \{g \in \mathcal{G} : \mathbb{P}(g(X, S) = i \mid Y = i, S = 0) = \mathbb{P}(g(X, S) = i \mid Y = i, S = 1), i = 0, 1\}.$$

In order to establish the form of such minimizer, we introduce the following assumption on the regression function.

Assumption 3.3.6 *For each $s \in \{0,1\}$ we require the mapping $t \in \mathbb{P}(\eta(X,S) \leq t \mid S = s)$ to be continuous on $(0,1)$, where for all $(x,s) \in \mathbb{R}^d \times \{0,1\}$, we let the regression function*

$$\eta(x,s) := \mathbb{P}(Y = 1 \mid X = x, S = s) = \mathbb{E}\left[Y \mid X = x, S = s\right]. \tag{3.3.14}$$

The following result establishes that the optimal *equalized odds* classifier is obtained recalibrating the Bayes classifier $g_B(X,S) = \mathbb{1}_{\{\eta(X,S) \geq 1/2\}}$, and its proof is included in the Appendix 3.6.3.

Proposition 3.3.7 (Optimal Rule) *Under Assumption 3.3.6, an optimal classifier g^* can be obtained for all $(x,s) \in \mathbb{R}^d \times \{0,1\}$ as*

$$g^*(x,1) = \mathbb{1}_{\{1 \leq 2\eta(X,1) - \theta_1^* \frac{\eta(X,1)}{\mathbb{P}(Y=1,S=1)} + \theta_0^* \frac{1-\eta(X,1)}{\mathbb{P}(Y=0,S=1)}\}}$$
$$g^*(x,0) = \mathbb{1}_{\{1 \leq 2\eta(X,0) + \theta_1^* \frac{\eta(X,0)}{\mathbb{P}(Y=1,S=0)} - \theta_0^* \frac{1-\eta(X,0)}{\mathbb{P}(Y=0,S=0)}\}},$$

where $(\theta_0^*, \theta_1^*) \in \mathbb{R}^2$ is determined from equations

$$\frac{\mathbb{E}_{X|S=1}\Big[\eta(X,1)g^*(X,1)\Big]}{\mathbb{P}(Y=1 \mid S=1)} = \frac{\mathbb{E}_{X|S=0}\Big[\eta(X,0)g^*(X,0)\Big]}{\mathbb{P}(Y=1 \mid S=0)}$$
$$\frac{\mathbb{E}_{X|S=1}\Big[(1-\eta(X,1))g^*(X,1)\Big]}{\mathbb{P}(Y=0 \mid S=1)} = \frac{\mathbb{E}_{X|S=0}\Big[(1-\eta(X,0))g^*(X,0)\Big]}{\mathbb{P}(Y=0 \mid S=0)}.$$

Remark 3.3.8 *Note that if $\theta_0^* = 0$ we recover the optimal fair equality of opportunity classifier in Chzhen et al. [2019]. If moreover $\theta_1^* = 0$ the above defined is the classical Bayes rule.*

We have quantified the cost with respect to the loss of the generalization error needed to ensure fairness in machine learning for classification and regression. If this price appears too high for the practitioner, the notion of fairness has to be weakened into a quantitative measure that can be adjusted for a trade-off between accuracy to the observations and fairness.

3.4 Quantifying fairness in machine learning

The importance of ensuring fairness in algorithmic outcomes has raised the need for designing procedures to remove the potential presence of bias. Yet building perfect fair models may lead to poor accuracy: changing the world into a fair one with positive action might decrease the efficiency defined as its similarity to the uses monitored through the test sample. While in some fields of application, it is desirable to ensure the highest possible level of fairness (see Shrestha and Yang [2019] for more details in applications of fair learning); in others, including Health Care or Criminal Justice, performance should not be decreased since the decisions would have serious implications for individuals and society. Hence, when perfect fairness requires to pay a too great price, resulting in poor generalization errors with respect to the unfair case, it is natural not to impose this strict condition but rather weaken the fairness constraint. In other words, it is of great interest to set a trade-off between fairness and accuracy, resulting in a relaxation of the notion of fairness that is frequently presented in the literature as *almost* or *approximate fairness*. To this aim, most methods approximate fairness desiderata through requirements on the lower order moments or other functions of distributions corresponding to different sensitive attributes.

From a procedural viewpoint, methods for imposing fairness are roughly divided in the literature into three families [Oneto and Chiappa, 2020, Dunkelau and Leuschel]. Methods in

the first family consist in pre-processing the data or in extracting representations that do not contain undesired biases (see e.g. Beutel et al. [2017], Calders et al. [2009], Calmon et al. [2017], Chierichetti et al. [2017], Edwards and Storkey [2015], Feldman [2015], Feldman et al. [2015], Fish and Lelkes [2015], Gordaliza et al. [2019], Johndrow and Lum [2019], Kamiran and Calders [2009, 2010, 2012], Madras et al. [2018a], Song et al. [2018], Zemel et al. [2013]), which can then be used as input to a standard machine learning model.

Methods in the second family, also referred to as in-processing, aim at enforcing a model to produce fair outputs through imposing fairness constraints into the learning mechanism. Some methods transform the constrained optimization problem via the method of Lagrange multipliers (see e.g. Agarwal et al. [2018], Berk et al. [2017a], Corbett-Davies et al. [2017], Cotter et al. [2018], Kearns et al. [2018], Narasimhan [2018], Zafar et al. [2017a, 2019]) or add penalties to the objective (see e.g. Bechavod and Ligett [2017], Donini et al. [2018], Dwork et al. [2018], Fukuchi et al. [2015], Hébert-Johnson et al. [2018], Kamiran et al. [2012], Kamishima et al. [2012], Kilbertus et al. [2017], Komiyama et al. [2018], Madras et al. [2018b], Mary et al. [2019], Nabi and Shpitser [2018], Narasimhan [2018], Oneto et al. [2019], Speicher et al. [2018], Yona and Rothblum [2018], Noroozi et al. [2019]), others use adversarial techniques to maximize the system ability to predict the target while minimizing the ability to predict the sensitive attribute [Zhang et al., 2018] and, finally, others rederive a new classifier from the first principles of distributional robustness that incorporates fairness criteria into a worst-case logarithmic loss minimization [Rezaei et al.].

Methods in the third family consist in post-processing the outputs of a model in order to make them fair (see e.g. Adler et al. [2018], Ali et al. [2019], Chzhen et al. [2019], Doherty et al. [2012], Feldman [2015], Fish et al. [2016], Hajian et al. [2012], Hardt et al. [2016], Kim et al. [2019], Kusner et al. [2017], Noriega-Campero et al. [2019], Pedreschi et al. [2009]).

As noticed in Oneto and Chiappa [2020], this grouping is imprecise and non exhaustive. Indeed, there are a number of works in the literature presenting alternative classifications, such as the survey Zhang and Liu [2020] that reviews existing literature on the fairness of data-driven sequential decision-making, which includes in practice most decision-making processes.

In the following we describe more deeply two different families of methods, which are non-mutually exclusive. First a group of in-processing methods which can be seen as a fair risk minimization problem and includes the majority of the contributions. On the other hand, a second category of methods based on optimal transport, which correspond mostly to pre or post processing approaches, since it is the preferred tool in this thesis for fair learning.

3.4.1 Fairness through Empirical Risk Minimization

We recall that the aim of a supervised machine learning algorithm is to learn the relationships between input characteristic variables and a target variable in order to forecast new observations. In the fair learning setting, we observe $(X_1, S_1, Y_1), \ldots, (X_n, S_n, Y_n)$ i.i.d observations drawn from an unknown distribution $\mathbb{P}$. Set the empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,S_i,Y_i}$. An almost-fair model will be obtained by minimizing the empirical risk

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i, S_i)),$$

with $\ell : (Y, \hat{Y}) \mapsto \ell(Y, \hat{Y}) \in \mathbb{R}^+$ a certain loss function measuring the quality of the prediction, and where the influence of the protected variable S in the forecast $\hat{Y}$ should be controlled. We note that such influence must be null in the case of perfect fairness and could be imposed by minimizing over a class $\mathcal{F}_{fair}$ satisfying certain stringent conditions. The classes $\mathcal{F}_{SP}$ or $\mathcal{F}_{EO}$, defined respectively in (3.3.1) and (3.3.2), are two possibilities for the minimization. In general,

a relaxation of the problem would enable control on the level of fairness of the learnt algorithm. This is proposed in the majority of the papers either by

(i) thresholding full-type fairness conditions, that is

$$\min_{f\in\mathcal{F}} R_n(f) \quad \text{such that } \delta(f(X,S),S,Y)\leq\varepsilon, \tag{3.4.1}$$

where δ is a measure of dependency (with $\delta(f(X,S),S,Y)=0$ in the perfect-fair case) and $\varepsilon>0$ represents the level of fairness; or

(ii) directly introducing the independence as a penalty into the objective

$$\min_{f\in\mathcal{F}} \{R_n(f)+\lambda\delta(f(X,S),S,Y)\}, \tag{3.4.2}$$

where $\lambda>0$ balances the contribution of both terms to get a trade-off between the bias and the efficiency of the algorithm.

Yet the main question becomes how to select the notion of independence measured above through the function δ. Several choices exist in the literature. According to the division of perfect fairness notions proposed in section 3.2.1, *almost fairness* requires quantifying the dependence between the distribution of the protected variable S and

(i) either the distribution of the forecast $\widehat{Y}$, or the conditional distribution of the forecast given the true value $\widehat{Y}|Y$,

(ii) or the expectation $\mathbb{E}\Phi(\widehat{Y})$ or $\mathbb{E}(\Phi(\widehat{Y})|Y)$, through a chosen function $\Phi:\mathbb{R}\to\mathbb{R}$.

Both points of view correspond to choices that can be made. In the following, we review how this framework summarizes most of the recent papers dealing with almost fairness.

3.4.1.1 Imposing conditions on the distributions

The first set of approaches to get fair predictive behaviour by adding constraints through conditions over the distributions has been studied in several papers. Depending on the basis of such conditions, the main proposals can be organised as follows:

(a) **Distance-based constraints**. According to the definition of fairness as independence criterion, this category of approaches aims at quantifying the distance between the probability distributions:

(i) $\mathcal{L}(\widehat{Y}|S=s)$, for all $s\in\mathcal{S}$; or $\mathcal{L}(\widehat{Y}\times S)$ and $\mathcal{L}(\widehat{Y})\times\mathcal{L}(S)$, if *statistical parity* is considered.

(ii) $\mathcal{L}(\widehat{Y}|Y,S=s)$, for all $s\in\mathcal{S}$, regarding to *equality of odds*.

The majority of the papers in this line of work considered Wasserstein distances and we summarize the main contributions hereafter. In Jiang et al. [2019] two different approaches to achieve *statistical parity* with Wasserstein-1 distance are proposed. First, a fast and practical approximation methodology to post-process the model outputs by enforcing the density functions of probabilities $\mathcal{L}(\hat{Y}\mid S=s)$ corresponding to groups of individuals with different sensitive attributes to coincide with their Wasserstein-1 barycenter distribution. Then, a penalization approach to binary logistic regression that aims at finding the model parameters minimizing the logistic loss under the constraint of small Wasserstein-1 distances between the empirical counterparts of measures $\mathcal{L}(\hat{Y}\mid S=s)$ and their empirical barycenter.

Wasserstein-type constraints for building fair classifiers has also been considered in Serrurier et al. [2019]. They provided algorithms which can incorporate both notions of fairness through 1-Wasserstein distance-based contraints. Yet sharing some similarities with Edwards and Storkey [2016], their approach is more flexible and enables to solve wider classes of fairness problems based on different adversarial architecture resulting in more suited loss functions. Neural networks are used to manage a large variety of input data structure (e.g. images) as well as output labels (multiclass, regression, images...). Their Wasserstein approximation using fairness benchmark datasets outperformed both classical fair algorithms (e.g fair SVM) as well as similar adversarial architectures based on Jensen or GAN losses (see references in the paper for more details.)

In Risser et al. [2019] algorithmic fairness is promoted by imposing closeness with respect to quadratic Wasserstein distance between the scores used to build an automatic decision rule. This regularization constraint is built with a deep neural network.

Specifically the concept of the barycenter in optimal transport theory is used in the recent paper Zehlike et al. [2020] to maximize decision maker utility under the chosen fairness constraints. They proposed the *Continuous Fairness Algorithm* which enables a continuous interpolation between different fairness definitions. This algorithm is able to handle cases of multi-dimensional discrimination of certain groups on grounds of several criteria. They included examples of credit applications, college admissions and insurance contracts; and mapped out the legal and policy implications of their approach.

(b) **Information theory-based contraints**. First contributions to this approach in the context of fair supervised learning started with the work of Kamishima et al. [2011], who designed an unfairness penalty term based on *statistical parity* criterion (referred to in their paper as *indirect prejudice*), which restricts the amount of mutual information between the prediction and the sensitive attribute. More precisely, they add a fairness regularization term in the objective function that penalizes the mutual information between the sensitive feature and the classifier decisions. In this way, this method treats the mutual information as the unfairness proxy. Their technique is only limited to the logistic regression classification model. Later in Kamishima et al. [2012] they used normalised MI to assess fairness in their *normalised prejudice index (NPI)*. Their focus is on binary classification with binary sensitive attributes, and the NPI is based on the independence fairness criterion. In such setting, mutual information is readily computable empirically from confusion matrices. This work is generalised in Fukuchi et al. [2015] for use in regression models by using a neutrality measure, which is shown to be equivalent to the independence criterion. They then use this neutrality measure to create inprocessing techniques for linear and logistic regression algorithms. Similarly, Ghassami et al. [2018] take an information theoretic approach to creating an optimisation algorithm that returns a predictor score that is fair with respect to the *equalized odds* criterion.

An information theory motivated framework is also proposed in Song et al. [2018] where the goal is to maximize what they called the *expressiveness* of representations of the data while satisfying certain fairness constraint. Expressiveness, as well as *statistical parity*, *equalized odds* and *equalized opportunity*, are expressed in terms of mutual information, and tractable upper and lower bounds of these mutual information objectives are obtained. A conexion between them and existing objectives such as maximum likelihood, adversarial training [Goodfellow et al., 2014], and variational autoencoders [Kingma and Welling, 2013, Rezende and Mohamed, 2015] is also presented. Their contribution serves as a unifying framework for existing work [Zemel et al., 2013, Edwards and Storkey, 2016, Madras et al., 2018a] on learning fair representations, being the first to provide direct user control over

the fairness of representations through fairness constraints that are interpretable by non-expert users.

In the regression setting, measuring group fairness criteria is computationally challenging, as it requires estimating information-theoretic divergences between conditional probability density functions. Recently Steinberg et al. [2020] introduced fast approximations of the *statistical parity*, *equality of odds* and *predictive parity* (there referred to as *independence*, *separation* and *sufficiency*, respectively; following ...) fairness criteria for regression models from their (conditional) mutual information definitions, and used such approximations as regularisers to enforce fairness within a regularised risk minimisation framework.

(c) **Kernel theory-based constraints**. Regularization is one of the key concepts in modern supervised learning, which allows imposing structural assumptions and inductive biases onto the problem at hand. It ranges from classical notions of sparsity, shrinkage, and model complexity to the more intricate regularization terms which allow building specific assumptions about the predictors into the objective functions, such as smoothness on manifolds [Belkin et al., 2006]. Such regularization viewpoint for algorithmic fairness was presented in Kamishima et al. [2012] in the context of classification, and was extended to regression and unsupervised dimensionality reduction problems with kernel methods in Pérez-Suay et al. [2017]. The latter falls within the framework of *statistical parity* and was the first work that considered this notion with continuous labels. They proposed kernel machines to exploit cross-covariance operators in Hilbert spaces. In particular, independence between predictor and sensitive variables is imposed by employing a kernel dependence measure, namely the Hilbert-Schmidt Independence Criterion (HSIC) Gretton et al. [2005], as a regularizer in the objective function.

Extentions of this work are presented in Li et al. [2019] where a general framework of empirical risk minimization with fairness regularizers and their interpretation is given. Secondly, they derived a Gaussian Process (GP) formulation of the fairness regularization framework, which allows uncertainty quantification and principled hyperparameter selection. Furthermore, they introduce a normalized version of the fairness regularizer which makes it less sensitive to the choice of kernel parameters. They demonstrate how the developed fairness regularization framework trades off model's predictive accuracy (with respect to potentially biased data) for independence to the sensitive covariates. It is worth noting that, in their setting, a function which produced the labels is not necessarily the function we wish to learn, so that the predictive accuracy is not necessarily a gold-standard criterion. Finally, we cite the work of [Tan et al., 2019] where the authors demonstrate the promise of learning a model-aware fair representation focusing on kernel-based models.

3.4.1.2 Imposing conditions on the expectation

On the other side, reinforcement of fair algorithmic behaviour has been also proposed by requiring conditions on the expected forecast in a large number of papers. More precisely, depending, on the one hand, on the desirable metric of fairness (as discussed in section 3.2); and on the other, on the nature of the target Y and the protected attribute S, the dependence measure δ is set out to control different kinds of indexes. We note that this control could be imposed following either (3.4.1) or (3.4.2).

1. For *statistical parity*, if $Y, S \in \{0, 1\}$, conditions on the probabilities of success across groups $\mathbb{P}(f(X, S) = 1|S)$ are considered, being the *mean difference score*

$$\mathbb{P}(f(X, S) = 1|S = 1) - \mathbb{P}(f(X, S) = 1|S = 0), \tag{3.4.3}$$

which was first introduced in Calders and Verwer [2010], and the disparate impact (3.3.6) the preferent choices in the literature. These are generalized to conditions on the expectation $\mathbb{E}(f(X,S)|S)$ (or $\mathbb{E}(\ell(f(X,S),Y)|S)$, with ℓ a loss function) or, from a sensitivity analysis point of view, on the variances $\mathrm{Var}(\mathbb{E}(f(X,S)|S))$ (or $\mathrm{Var}(\mathbb{E}(\ell(f(X,S),Y)|S))$).

2. For *equality of odds*, if $Y, S \in \{0,1\}$, then the goal is similar as before but taking into account the true values of the target Y. Namely, the differences between TPR and FPR, that is

$$\mathbb{P}(f(X,S)=i|Y=i,S=1) - \mathbb{P}(f(X,S)=i|Y=i,S=0), \text{ for } i=0,1, \tag{3.4.4}$$

are usually considered. Besides, in a less demanding way, others focus on the difference between the overall accuracies

$$\mathbb{P}(Y \neq \hat{Y} \mid S=0) - \mathbb{P}(Y \neq \hat{Y} \mid S=1). \tag{3.4.5}$$

In a wider setup, this is extended to conditions on the expectation $\mathbb{E}(f(X,S)|Y,S)$ (or $\mathbb{E}(\ell(f(X,S),Y)|Y,S)$) or the variance $\mathrm{Var}(\mathbb{E}(f(X,S)|Y,S)$ (or $\mathrm{Var}(\mathbb{E}(\ell(f(X,S),Y)|Y,S))$.

Given this overview summarizing the majority of proposals for relaxing the notion of fairness through conditions on the input and output distributions of the algorithm, we cite some of the main contributions to this approach. One of the first was the work of Zemel et al. [2013] which, based on Dwork et al. [2012], combined pre-processing and inprocessing by jointly learning a 'fair' representation of the data and the classifier parameters. The joint representation is learnt using a multi-objective loss function that ensures that (i) the resulting representations do not lead to disparate impact, (ii) the reconstruction loss from the original data and intermediate representations is small and (iii) the class label can be predicted with high accuracy. This approach has two main limitations: i) it leads to a non-convex optimization problem and does not guarantee optimality, and ii) the accuracy of the classifier depends on the dimension of the fair representation, which needs to be chosen rather arbitrarily. Inspired by Zemel et al. [2013], the methods of Edwards and Storkey [2016] and Madras et al. [2018a] also aim at learning fair representations of the data.

In Zafar et al. [2017a] methods for training decision boundary-based classifiers without *disparate mistreatment* (recall (3.2.7)) are described, with further extensions to existing notions *disparate treatment* and *disparate impact* in Zafar et al. [2017b]. Their proposals, as well as the results of several experiments and applications to well-known real datasets, have been collected later in Zafar et al. [2019]. They noticed that taking in the above formulation (3.4.1) the dependence measure δ in terms of the accuracies in (3.4.5), and similarly for (3.4.4), ensures that the classifier chooses the optimal decision boundary within the space of fair boundaries specified by the constraints but yields to a very challenging problem. The reason is two-fold: first, the fairness constraints lead to non-convex formulations; and second, the probabilities defining such constraints are function having saddle points, which further complicates the procedure for solving non-convex optimization problems [Dauphin et al., 2014]. Therefore, they proposed a relaxation of these (non-convex) fairness constraints into proxy conditions, each in the form of a convex-concave (or, difference of convex) function using a covariance measure of decision boundary fairness. They design fair logistic regression classifiers and linear and nonlinear SVMs as examples and heuristically solve the resulting optimization problem for a convex loss function. Adding constraints to the classification model is also in the line of work of Goh et al. [2016], Woodworth et al. [2017] and Quadrianto and Sharmanska [2017]. While the constraints are similar to those in Zafar et al. [2019], the first two are only limited to a single specific loss function and the third one to a single notion of unfairness.

Another approach in pursuit of fairness as *equality of odds* in binary classifiers learned over individuals from two populations is presented in Bechavod and Ligett [2017]. They validate the ability of such approach to achieve both fairness and high accuracy, implementing and testing it on multiple datasets from the fields of criminal risk assessment, credit, lending, and college admissions. Later in Agarwal et al. [2018] both *statistical parity* and *equalized odds* conditions are viewed as a special case of a general set of linear constraints. Based on that, the minimization problem is shown to be reduced to a sequence of cost-sensitive classification problems, whose solutions yield a randomized classifier with the lowest (empirical) error subject to the desired constraints.

In Menon and Williamson [2018] *disparate impact* and *mean difference* indexes are related to cost-sensitive risks and the tradeoffs between performance of in the problem of learning with these fairness constraint are studied. They showed that the optimal classifier for these cost-sensitive measures is an instance-dependent thresholding of the classprobability function, and quantify the degradation in performance by a measure of alignment of the target and sensitive variable. They also use such analysis to derive a simple plugin approach for the fairness problem. Finally, in the classification setting we metion also Kearns et al. [2018], who considered the problem of learning binary classifiers subject to *equal opportunity* and *statistical parity* constraints when the number of protected groups is large.

In the fair regression framework, Zafar et al. [2017a] suggested a relaxed notion of non-discrimination based on first order moments

$$\mathbb{E}(\hat{Y}|Y = y, S = 0) = \mathbb{E}(\hat{Y}|Y = y, S = 1)$$

and proposed optimizing a convex loss subject to an approximation of this constraint. With a similar aim, in previously cited paper Woodworth et al. [2017] (see section 3.3.2.1) they proposed a relaxation of the criterion of *equalized odds* by a more tractable notion of non-discrimination based on second order moments. In particular, they proposed the notion of *equalized correlations*. Later, in Agarwal et al. [2019] the fair regression problem is studied in a predictive setting where $\mathcal{X}$ could be continuous and high-dimesional, $\mathcal{S}$ is discrete, and $\mathcal{Y} \subseteq [0, 1]$ could be discrete (but embedded in $[0, 1]$) or continuous. Two different constraints in the minimization (3.4.1) are considered in this work. Firstly, a relaxation of *statistical parity* is proposed as, for all $z \in [0, 1]$ and all $s \in \mathcal{S}$,

$$|\mathbb{P}(f(X) \geq z \mid S = s) - \mathbb{P}(f(X) \geq z| \leq \varepsilon_s, \tag{3.4.6}$$

where the slack $\varepsilon_s > 0$ bounds the allowed departure of the CDF of $f(X)$ conditional on $S = s$ from the CDF of $f(X)$. Note that the protected variable S is not explicitely considered as input. The difference between CDFs is measured in the $\infty-$norm corresponding to the Kolmogorov-Smirnov statistic. On the other hand, they also propose to guarantee fairness through the criteria *bounded group loss*

$$\mathbb{E}(\ell(f(X), Y)|S) \leq \varepsilon_s \tag{3.4.7}$$

where, in fact, the threshold is uniform for all the classes in the definition, but, for the sake of flexibility, it is allowed to specify different bounds $\varepsilon_s > 0$ for each attribute value in the loss minimization. Hence, fair regression with *bounded group loss* minimizes the overall loss, while controlling the worst loss on any protected group. By Lagrangian duality, this is equivalent to minimizing the worst loss on any group while maintaining good overall loss (referred to as *max-min fairness*). Unlike *overall accuracy equality* in classification Dieterich et al. [2016], which requires the losses on all groups to be equal, they claimed that *bounded group loss* does not force an artificial decrease in performance on every group just to match the hardest-to-predict group. They also generalized their approach to randomized predictors to achieve better fairness-accuracy trade-off.

We finally cite the recent algorithm in Oneto and Chiappa [2020] called *General Fair Empirical Risk Minimization (G-FERM)* that generalizes the *Fair Empirical Risk Minimization* approach introduced in Donini et al. [2018]. In this work, they also specify the method for the case in which the underlying space of models is a RKHS and show how the in-processing G-FERM approach described above can be translated into a pre-processing approach.

3.4.2 Fairness through Optimal Transport

Most methods obtain fair models by imposing approximations of fairness desiderata through constraints on lower order moments or other functions of distributions corresponding to different sensitive attributes (this is also what most popular fairness definitions require). As observed in Oneto and Chiappa [2020], whilst facilitating model design, not imposing constraints on the full shapes of relevant distributions can be problematic. One existing approach that does work this way proposes to match distributions corresponding to different sensitive attributes either in the space of model outputs or in the space of model inputs (or latent representations of the inputs) using optimal transport theory, which correspond to post and pre-processing methods, respectively. We note that the in-processing methods based on optimal transport are those imposing constraints in terms of the Wasserstein distance and have already been described above (see in section 3.4.1.1(a)).

The idea of the pre-processing based methods to obtain fair treatment consists in blurring the value of the protected class by transporting the original distribution of the input, conditionally to this value, towards their Wasserstein's barycenter. It was first considered in the binaty classification problem in Feldman et al. [2015], Johndrow and Lum [2019] or Hacker and Wiedemann [2017], and later improved in Gordaliza et al. [2019]. In this work, the choice of the weighted Wasserstein's barycenter with respect to the weights of the protected clased is formally justified (see Theorem 4.3.3.) in terms of the minimal excess risk when considering the classfier trained from the repaired data. Moreover, they propose to set an accuracy-fairness trade-off through a partial repair approach called *random repair*, which it is shown to outperform the previous geometric repair in Feldman et al. [2015].

The work in Chiappa et al. [2020], Jiang et al. [2019] presents an approach to fair classification and regression that is applicable to many fairness criteria. In particular, they introduce the notion of *Strong Demographic Parity*, which extends the *statistical parity* to a fair multi-classification and regression problem. Based on that, in Oneto and Chiappa [2020] they derived a simple post-processing method withing this framework to achieve *Strong Demographic Parity* by transporting distributions to their Wasserstein barycenter. They also propose a partial transportation for setting a fairness-accuracy trade-off called the *Wasserstein 2-Geodesic* method.

3.5 Conclusions

In this paper, we have presented a review of mathematical models designed to handle the issue of bias in machine learning. Due to the large number of definitions, we have proposed a probabilistic framework to understand the relationships between fairness and the notion of independence or conditional independence. Hence imposing fairness is here modeled as imposing independence with respect to the sensitive variable and constraints are naturally driven by the choice of different measures for this independence. Within this framework, it becomes thus possible to give another insight at several notions of fairness and also to quantity their effect on the decision rule. In particular, we can defined and then compute in some cases the so-called price for fairness to quantify the real impact of fairness constraint on the behavior of a machine learning algorithm. This study provides a better understanding of fair learning, each different definition of fairness

leading to different behaviors that can be compared in some cases. Yet many cases remain open to further research to obtain a full theoretical framework of fair learning.

Moreover, we point out that we did not consider in this study many new interesting points of view on fairness that deserve a specific study. In very particular, understanding fairness from a causal point of view or using counter-examples as in Loftus et al. [2018] and Kusner et al. [2017] or Black et al. [2020] could provide another interpretation for fairness in machine learning.

3.6 Appendix to Chapter 3

3.6.1 Proofs of section 3.2.3

Proof of Proposition 3.2.1. Observe that if $S \perp\!\!\!\perp \hat{Y}$ and $\hat{Y} \perp\!\!\!\perp Y \mid S$ then either $S \perp\!\!\!\perp Y$ or $\hat{Y} \perp\!\!\!\perp Y$.

□

Proof of Proposition 3.2.2. It suffices to observe that if $S \not\perp\!\!\!\perp Y$ and $S \perp\!\!\!\perp Y \mid \hat{Y}$ then $S \not\perp\!\!\!\perp \hat{Y}$.
□

Proof of Proposition 3.2.3. $S \perp\!\!\!\perp \hat{Y} \mid Y$ and $S \perp\!\!\!\perp Y \mid \hat{Y}$ implies $S \perp\!\!\!\perp (\hat{Y}, Y)$, and then $S \perp\!\!\!\perp Y$.
□

3.6.2 Proofs of section 3.3.2.1

We start recalling some facts about Gaussian random variables.

Proposition 3.6.1 *If (U, V, W) are jointly Gaussian, then*

- *Conditional expectation $\mathbb{E}(U|V)$ is linear in V and is given by*

$$\mathbb{E}(U|V) = \mathbb{E}(U) + \Sigma_{U,V}\Sigma_V^{-1}(V - \mathbb{E}(V))$$

- *Conditional covariance $\Sigma_{(U,V)|W}$ does not depend on W and is given by*

$$\Sigma_{(U,V)|W} = \Sigma_{U,V} - \Sigma_{U,W}\Sigma_W^{-1}\Sigma_{U,W}^T$$

Proof of Proposition 3.3.5. In the particular normal model, this independence means that the elements in positions $(1, 2)$ and $(2, 1)$ of the covariance matrix of random vector $(g(X, S), S \mid Y)$ are exactly zero. Therefore, the class of fair predictors is written as

$$\mathcal{F}_{EO} := \{g : \mathcal{X} \times \mathcal{S} \to R^d : Cov(g(X, S), S \mid Y) = 0\} \tag{3.6.1}$$

More precisely, previous condition can be written in terms of the covariances of (X, S, Y) and the coefficients (β_0, β) of the linear model (3.3.9). Observe that the joint distribution of the random vector $(g_{\beta_0,\beta}(X, S), S, Y)$ is

$$\begin{bmatrix} \beta_0 S + \beta^T X \\ S \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \beta_0\mu_S + \beta^T\mu_X \\ \mu_S \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_Y \end{bmatrix} \right),$$

where

$$\Sigma_1 = \begin{bmatrix} \beta_0^2\Sigma_S + \beta^T\Sigma_X\beta + 2\beta_0\beta^T\Sigma_{XS} & \beta_0\Sigma_S + \beta^T\Sigma_{XS} \\ \beta_0\Sigma_{SY} + \beta^T\Sigma_{XY} & \Sigma_S \end{bmatrix} \in \mathbb{R}^{2\times 2},$$
$$\Sigma_{12} = \begin{bmatrix} \beta_0\Sigma_{SY} + \beta^T\Sigma_{XY} \\ \Sigma_{SY} \end{bmatrix} \in \mathbb{R}^{2\times 1}.$$

Hence, from Proposition 3.6.1, we know that

$$Cov(g(X,S), S \mid Y) = \Sigma_1 - \frac{1}{\Sigma_Y}\Sigma_{12}\Sigma_{12}^T.$$

Substituting the expressions above for Σ_1 and Σ_2, we obtain that $g_{\beta_0,\beta} \in \mathcal{F}_{EO}$ if and only if

$$(\beta_0\Sigma_S + \beta^T\Sigma_{XS})\Sigma_Y = \Sigma_{SY}(\beta_0\Sigma_{SY} + \beta^T\Sigma_{XY}).$$

Then the optimal EO-fair predictor in this setting is the solution to the following optimization problem:

$$\left(\hat{\beta}_{0,fair}, \hat{\beta}_{fair}\right) := argmin_{(\beta_0,\beta)\in\mathcal{F}_{EO}}\mathbb{E}\left[(Y - g_{\beta_0,\beta}(X,S))^2\right] \tag{3.6.2}$$
$$\mathcal{F}_{EO} = \{(\beta_0,\beta) \in \mathbb{R}\times\mathbb{R}^p \text{ such that } \beta^T\left(\Sigma_{XS}\Sigma_Y - \Sigma_{SY}\Sigma_{XY}\right) + \beta_0\left(\Sigma_S\Sigma_Y - \Sigma_{SY}^2\right) = 0\}.$$

We note that Cauchy-Schwarz inequality together with the assumption that Y and S are not linearly dependent ensure $\Sigma_S\Sigma_Y - \Sigma_{SY}^2 > 0$. Then we obtain that the class of EO-fair predictors $(\beta_0,\beta) \in \mathcal{F}_{EO}$ are such that $\beta_0 = \beta^T C_{S,X,Y}$, where

$$C_{S,X,Y} := \left(\frac{\Sigma_{XS}\Sigma_Y - \Sigma_{SY}\Sigma_{XY}}{\Sigma_S\Sigma_Y - \Sigma_{SY}^2}\right) \in \mathbb{R}^{p\times 1}.$$

Hence, the optimal EO-fair predictor (3.6.2) can be obtained equivalently

$$\hat{\beta}_{fair} = argmin_{\beta\in\mathbb{R}^p}\mathbb{E}\left[\left(Y - \beta^T(X + SC_{S,X,Y})\right)^2\right].$$

Now if we denote $Z := X + SC_{S,X,Y}$, it is easy to check that the optimal EO-fair predictor can be exactly computed as

$$\begin{aligned} \hat{\beta}_{fair} &= \Sigma_Z^{-1}\Sigma_{Z,Y}, \text{ where} \\ \Sigma_Z &= \Sigma_X + \Sigma_S C_{S,X,Y}C_{S,X,Y}^T + C_{S,X,Y}\Sigma_{XS}^T + \Sigma_{XS}C_{S,X,Y}^T \\ \Sigma_{ZY} &= \Sigma_{XY} + \Sigma_{SY}C_{S,X,Y}. \end{aligned}$$

□

3.6.3 Proofs of section 3.3.2.2

Proof of Proposition 3.3.7. Let us consider the following minimization problem

$$(*) := \min_{g\in\mathcal{G}}\{\mathcal{R}(g) : \mathbb{P}(g(X,S) = i \mid Y = i, S = 1) = \mathbb{P}(g(X,S) = i \mid Y = i, S = 0), i = 0, 1\}.$$

Using the weak duality we can write

$$
\begin{aligned}
(*) =& \\
=& \min_{g\in\mathcal{G}} \max_{(\lambda_0,\lambda_1)\in\mathbb{R}^2} \left\{ \mathcal{R}(g) + \sum_{i=0,1} \lambda_i \left[\mathbb{P}(g(X,S)=i \mid Y=i, S=1) - \mathbb{P}(g(X,S)=i \mid Y=i, S=0) \right] \right\} \\
\geq& \max_{(\lambda_0,\lambda_1)\in\mathbb{R}^2} \min_{g\in\mathcal{G}} \left\{ \mathcal{R}(g) + \sum_{i=0,1} \lambda_i \left[\mathbb{P}(g(X,S)=i \mid Y=i, S=1) - \mathbb{P}(g(X,S)=i \mid Y=i, S=0) \right] \right\} \\
=:& (**).
\end{aligned}
$$

We first study the objective function of the max min problem $(**)$, which is equal to

$$
\mathbb{P}(g(X,S)\neq Y) + \sum_{i=0,1} \lambda_i \left(\mathbb{P}(g(X,S)=i \mid Y=i, S=1) - \mathbb{P}(g(X,S)=i \mid Y=i, S=0) \right).
$$

The first step of the proof is to simplify the expression above to linear functional of the classifier g. Notice that we can write for the first term

$$
\begin{aligned}
\mathbb{P}(g(X,S)\neq Y) =& \mathbb{P}(g(X,S)=0, Y=1) + \mathbb{P}(g(X,S)=1, Y=0) \\
=& \mathbb{P}(g(X,S)=1) + \mathbb{P}(Y=1) - \mathbb{P}(g(X,S)=1, Y=1) - \mathbb{P}(g(X,S)=1, Y=1) \\
=& \mathbb{P}(g(X,S)=1) + \mathbb{P}(Y=1) - 2\mathbb{P}(g(X,S)=1, Y=1) \\
=& \mathbb{P}(Y=1) + \mathbb{E}\left[g(X,S)\right] - 2\mathbb{P}(S=1)\mathbb{E}\left[\mathbb{1}_{g(X,S)=1,Y=1} \mid S=1\right] \\
& - 2\mathbb{P}(S=0)\mathbb{E}\left[\mathbb{1}_{g(X,S)=1,Y=1} \mid S=0\right] \\
=& \mathbb{P}(Y=1) - \mathbb{P}(S=1)\mathbb{E}_{X|S=1}\left[g(X,1)(2\eta(X,1)-1)\right] \\
& - \mathbb{P}(S=0)\mathbb{E}_{X|S=0}\left[g(X,0)(2\eta(X,0)-1)\right].
\end{aligned}
$$

Moreover, for $s=0,1$, we can write for the rest four terms in the objetive function

$$
\begin{aligned}
&\mathbb{P}(g(X,S)=1 \mid Y=1, S=s) = \frac{\mathbb{P}(g(X,S)=1, Y=1 \mid S=s)}{\mathbb{P}(Y=1 \mid S=s)} = \frac{\mathbb{E}_{X|S=s}\left[g(X,s)\eta(X,s)\right]}{\mathbb{P}(Y=1 \mid S=s)} \\
&\mathbb{P}(g(X,S)=0 \mid Y=0, S=s) = 1 - \mathbb{P}(g(X,S)=1 \mid Y=0, S=s) \\
&= 1 - \frac{\mathbb{E}_{X|S=s}\left[g(X,s)(1-\eta(X,s))\right]}{\mathbb{P}(Y=0 \mid S=s)}.
\end{aligned}
$$

Using these, the objective of $(**)$ can be simplified as

$$
\begin{aligned}
\mathbb{P}(Y=1) + \mathbb{E}_{X|S=1}\Big[g(X,1) \Big(& \eta(X,1) \left(\frac{\lambda_1}{\mathbb{P}(Y=1 \mid S=1)} + \frac{\lambda_0}{1-\mathbb{P}(Y=1 \mid S=1)} - 2\mathbb{P}(S=1) \right) \\
& + \mathbb{P}(S=1) - \frac{\lambda_0}{1-\mathbb{P}(Y=1 \mid S=1)} \Big) \Big] \\
+ \mathbb{E}_{X|S=0}\Big[g(X,0) \Big(& \eta(X,0) \left(-\frac{\lambda_1}{\mathbb{P}(Y=1 \mid S=0)} - \frac{\lambda_0}{1-\mathbb{P}(Y=1 \mid S=0)} - 2\mathbb{P}(S=0) \right) \\
& + \mathbb{P}(S=0) + \frac{\lambda_0}{1-\mathbb{P}(Y=1 \mid S=0)} \Big) \Big].
\end{aligned}
$$

For every $\lambda := (\lambda_0, \lambda_1) \in \mathbb{R}^2$ a minimizer g^*_λ of the problem $(**)$ can be written for all $x \in \mathbb{R}^d$ as

$$\begin{aligned}
g^*_\lambda(x,1) &= \mathbb{1}_{\{\eta(X,1)\left(\frac{\lambda_1}{\mathbb{P}(Y=1|S=1)}+\frac{\lambda_0}{1-\mathbb{P}(Y=1|S=1)}-2\mathbb{P}(S=1)\right)+\mathbb{P}(S=1)-\frac{\lambda_0}{1-\mathbb{P}(Y=1|S=1)}\le 0\}} \\
&= \mathbb{1}_{\{1-\eta(X,1)\left(2-\frac{\lambda_1}{\mathbb{P}(Y=1,S=1)}-\frac{\lambda_0}{\mathbb{P}(Y=0,S=1)}\right)-\frac{\lambda_0}{\mathbb{P}(Y=0,S=1)}\le 0\}} \\
&= \mathbb{1}_{\{1\le 2\eta(X,1)-\lambda_1\frac{\eta(X,1)}{\mathbb{P}(Y=1,S=1)}+\lambda_0\frac{1-\eta(X,1)}{\mathbb{P}(Y=0,S=1)}\}} \\
g^*_\lambda(x,0) &= \mathbb{1}_{\{\eta(X,0)\left(-\frac{\lambda_1}{\mathbb{P}(Y=1|S=0)}-\frac{\lambda_0}{1-\mathbb{P}(Y=1|S=0)}-2\mathbb{P}(S=0)\right)+\mathbb{P}(S=0)+\frac{\lambda_0}{1-\mathbb{P}(Y=1|S=0)}\le 0\}} \\
&= \mathbb{1}_{\{1-\eta(X,0)\left(2+\frac{\lambda_1}{\mathbb{P}(Y=1,S=0)}+\frac{\lambda_0}{\mathbb{P}(Y=0,S=0)}\right)+\frac{\lambda_0}{\mathbb{P}(Y=0,S=0)}\le 0\}} \\
&= \mathbb{1}_{\{1\le 2\eta(X,0)+\lambda_1\frac{\eta(X,0)}{\mathbb{P}(Y=1,S=0)}-\lambda_0\frac{1-\eta(X,0)}{\mathbb{P}(Y=0,S=0)}\}}.
\end{aligned}$$

It is interesting to observe that for $\lambda_0 = 0$ we recover the optimal equal opportunity classifier obtained first in Chzhen et al. [2019]. If in addition $\lambda_1 = 0$, then we recover the Bayes classifier. Now, substituting this classifier into the objective of $(**)$ we arrive at

$$\begin{aligned}
\mathbb{P}(Y=1) - \min_{(\lambda_0,\lambda_1)\in\mathbb{R}^2} \Big\{ &\mathbb{E}_{X|S=1}\Big[\eta(X,1)\left(-2\mathbb{P}(S=1)+\frac{\lambda_1}{\mathbb{P}(Y=1\mid S=1)}+\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=1)}\right) \\
&\qquad +\mathbb{P}(S=1)-\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=1)}\Big]_+ \\
&+\mathbb{E}_{X|S=0}\Big[\eta(X,0)\left(-2\mathbb{P}(S=0)-\frac{\lambda_1}{\mathbb{P}(Y=1\mid S=0)}-\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=0)}\right) \\
&\qquad +\mathbb{P}(S=0)+\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=0)}\Big]_+\Big\}.
\end{aligned}$$

We observe that the mappings

$$\begin{aligned}
(\lambda_0,\lambda_1) \mapsto\ &\mathbb{E}_{X|S=1}\Big[\eta(X,1)\left(-2\mathbb{P}(S=1)+\frac{\lambda_1}{\mathbb{P}(Y=1\mid S=1)}+\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=1)}\right) \\
&\qquad +\mathbb{P}(S=1)-\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=1)}\Big]_+ \\
(\lambda_0,\lambda_1) \mapsto\ &\mathbb{E}_{X|S=0}\Big[\eta(X,0)\left(-2\mathbb{P}(S=0)-\frac{\lambda_1}{\mathbb{P}(Y=1\mid S=0)}-\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=0)}\right) \\
&\qquad +\mathbb{P}(S=0)+\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=0)}\Big]_+
\end{aligned}$$

are convex, therefore we can write the first order optimality conditions as

$$\begin{aligned}
\mathbf{0} \in\ &\partial_\lambda\mathbb{E}_{X|S=1}\Big[\eta(X,1)\left(-2\mathbb{P}(S=1)+\frac{\lambda_1}{\mathbb{P}(Y=1\mid S=1)}-\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=1)}\right) \\
&\qquad +\mathbb{P}(S=1)-\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=1)}\Big]_+ \\
&+\partial_\lambda\mathbb{E}_{X|S=0}\Big[\eta(X,0)\left(-2\mathbb{P}(S=0)-\frac{\lambda_1}{\mathbb{P}(Y=1\mid S=0)}-\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=0)}\right) \\
&\qquad +\mathbb{P}(S=0)+\frac{\lambda_0}{1-\mathbb{P}(Y=1\mid S=0)}\Big]_+
\end{aligned}$$

Under Assumption 3.3.6 this subgradient is reduced to the gradient almost surely, thus we have

the following two conditions on the optimal value of λ^*

$$\frac{\mathbb{E}_{X|S=1}\Big[\eta(X,1)g^*_{\lambda^*}(X,1)\Big]}{\mathbb{P}(Y=1 \mid S=1)} = \frac{\mathbb{E}_{X|S=0}\Big[\eta(X,0)g^*_{\lambda^*}(X,0)\Big]}{\mathbb{P}(Y=1 \mid S=0)} \tag{3.6.3}$$

$$\frac{\mathbb{E}_{X|S=1}\Big[(1-\eta(X,1))g^*_{\lambda^*}(X,1)\Big]}{\mathbb{P}(Y=0 \mid S=1)} = \frac{\mathbb{E}_{X|S=0}\Big[(1-\eta(X,0))g^*_{\lambda^*}(X,0)\Big]}{\mathbb{P}(Y=0 \mid S=0)} \tag{3.6.4}$$

and the pair $(\lambda^*, g^*_{\lambda^*})$ is a solution of the dual problem $(**)$. By the definition of the regression function (3.3.14), we note that previous conditions (3.6.3) and (3.6.4) can be written as

$$\begin{aligned}
&\mathbb{P}(g^*_{\lambda^*}(X,1)=1 \mid Y=1, S=1) = \mathbb{P}(g^*_{\lambda^*}(X,0)=1 \mid Y=1, S=0)\\
&\mathbb{P}(g^*_{\lambda^*}(X,1)=1 \mid Y=0, S=1) = \mathbb{P}(g^*_{\lambda^*}(X,1)=1 \mid Y=0, S=0)
\end{aligned}$$

which implies that the classifier $g^*_{\lambda^*} \in \mathcal{F}_{EO}$, that is, it is fair in the EO sense.

Finally, it remains to show that $g^*_{\lambda^*}$ is actually an optimal classifier. Indeed, since $g^*_{\lambda^*}$ is fair we can write on the one hand

$$\mathcal{R}(g^*_{\lambda^*}) \geq \min_{g\in\mathcal{G}}\{\mathcal{R}(g) : \mathbb{P}(g(X,S)=i \mid Y=i, S=0) = \mathbb{P}(g(X,S)=i \mid Y=i, S=1), i=0,1\} = (*).$$

On the other hand, the pair $(\lambda^*, g^*_{\lambda^*})$ is a solution of the dual problem $(**)$, thus we have

$$\begin{aligned}
(*) &\geq \mathcal{R}(g^*_{\lambda^*}) + \sum_{i=0,1} \lambda^*_i \left(\mathbb{P}(g^*_{\lambda^*}(X,S)=i \mid Y=i, S=0) - \mathbb{P}(g^*_{\lambda^*}(X,S)=i \mid Y=i, S=1))\right\}\\
&= \mathcal{R}(g^*_{\lambda^*}).
\end{aligned}$$

It implies that the classifier $g^*_{\lambda^*}$ is optimal, hence $g^* \equiv g^*_{\lambda^*}$.

□

Chapter 4

Obtaining Fairness using Optimal Transport Theory

The content of this chapter has been presented at the *International Conference of Machine Learning (Los Angeles, june 2019)* and it is published in the book of Proceedings of Machine Learning Research as Gordaliza et al. [2019].

Contents

In the fair classification setup, we recast the links between fairness and predictability in terms of probability metrics. We analyze repair methods based on mapping conditional distributions to the Wasserstein barycenter. We propose a *Random Repair* which yields a tradeoff between minimal information loss and a certain amount of fairness.

4.1 Introduction

Along the last decade, machine learning methods have become more popular to build decision algorithms. Originally meant for Internet recommendation systems, they are now widely used in a large number of very sensitive areas such as medicine, human ressources with hiring policies,

banking and insurance (lending), police and justice with criminal sentencing, see for instance Berk et al. [2017b], Pedreschi et al. [2012] or Friedler et al. [2019]. The decisions made by what is now referred to as AI have a growing impact on human life. The whole machinery of these techniques relies on the fact that a decision rule can be learnt by looking at a subset of labeled examples, the learning sample, and then is applied to the whole population which is assumed to follow the same underlying distribution. So the decision is highly influenced by the choice of the learning set.

In some cases, this learning sample may present some bias or discrimination that could possibly be learnt by the algorithm and then propagated to the entire population through automatic decisions, providing a mathematical legitimacy for this unfair treatment. When giving algorithms the power to make automatic decisions, the danger may come that the reality may be shaped according to their prediction, thus reinforcing their beliefs in the model which is learnt. Hence, achieving fair treatment is one of the growing fields of interest in machine learning. For a recent survey on this topic we refer to Zafar et al. [2017a] or Friedler et al. [2019].

Classification algorithms are one particular focus of fairness concerns since classifiers map individuals to outcomes. Some variables, such as sex, age or ethnic origin, are potentially sources of unfair treatment since they enable to create information that should not be processed out by the algorithm. Such variables are called in the literature protected variables. An algorithm is said to be fair with respect to these attributes when its outcome does not allow to make inference on the information they convey. Of course, the naive solution of ignoring these attributes when learning the classifier does not ensure this, since the protected variables may be closely correlated with other features enabling a classifier to reconstruct them.

Two solutions have been considered in the fair learning literature. The first one consists in changing the classifier in order to make it not correlated to the protected attribute. We refer for instance to Zafar et al. [2017a], Bechavod and Ligett [2017] or Donini et al. [2018]. Yet, explaining how the classifier is chosen may be seen too intrusive for many companies, or some of them may not even be able to change the way they build their models. Hence, a second solution consists in modifying the input data so that predictability of the protected attribute is impossible, whatever the classifier we train. The idea consists in blurring the value of the protected class trying to obtain a fair treatment. This point of view has been proposed in Feldman et al. [2015], Johndrow and Lum [2019] and Hacker and Wiedemann [2017], for instance.

In this paper, we first provide in Section 4.2 a statistical analysis of the Disparate Impact definition and recast some of the ideas developed in Feldman et al. [2015] to stress the links between fairness, predictability and the distance between the distributions of the variables given the protected attribute. Then, in Section 4.3 we provide first in 4.3.1 some theoretical justifications of the methodology proposed by previous authors (for one-dimensional data) to blur the data using the barycenter of the conditional distribution with respect to the Wasserstein distance. These methods are called either *total* or *partial repair*. Then in Section 4.3.2, we propose another methodology called *random repair* to transform the data in order to achieve a tradeoff between a minimal information loss of the classification task and still a certain level of fairness. We extend in Section 4.4 this procedure to the multidimensional case and provide a feasible algorithm to achieve the repair using the notion of Wasserstein barycenter. Finally application to simulated data in Section 4.5 enables to study the efficiency of the proposed procedures.

4.2 Framework for the fairness problem

Consider the probability space $(\Omega, \mathcal{B}, \mathbb{P})$, with $\mathcal{B}$ the Borel $\sigma-$algebra of subsets of $\mathbb{R}^d$ and $d \geq 1$. In this paper, we tackle the problem of forecasting a binary variable $Y : \Omega \to \{0,1\}$, using observed covariates $X : \Omega \to \mathbb{R}^d,\ d \geq 1$. We assume moreover that the population can be divided into two categories that represent a bias, modeled by a variable $S : \Omega \to \{0,1\}$. This variable is called the protected attribute and takes the values $S = 0$ for the *minority* (assumed to be the unfavored class), and $S = 1$ for the *default* (and, usually, favored class). We also introduce also a notion of positive prediction: $Y = 1$ represents a *success* while $Y = 0$ is a *failure*. Hence, the classification problem aims at predicting a success from variables X, using a family $\mathcal{G}$ of binary classifiers $g : \mathbb{R}^d \to \{0,1\}$. For every $g \in \mathcal{G}$, the outcome of the classification will be the prediction $\hat{Y} = g(X)$. We refer to Bousquet et al. [2004] for a complete description of classification problems in statistical learning.

In this framework, discrimination or unfairness of the classification procedures, appears as soon as the prediction and the protected attribute are too closely related, in the sense that statistical inference on Y may lead to learn the distribution of the protected attribute S. This issue has received lots of attention in the last years and several ways to quantify this *discrimination bias* have been given. We refer for instance to Lum and Johndrow [2016], Chouldechova [2017] or Bechavod and Ligett [2017] for the analysis of fairness in machine learning. Here we focus on the definition given in Feldman et al. [2015] or Berk et al. [2017b]. A classifier $g : \mathbb{R}^d \to \{0,1\}$ is said to achieve *statistical parity*, with respect to the joint distribution of (X, S), if

$$\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1). \tag{4.2.1}$$

This means that the probability of a successful outcome is the same across the groups. Yet, the independence described in (5.1.5) is difficult to achieve and may not exist in real data. An index called *disparate impact* (DI) of the classifier g with respect to (X, S) has been introduced in Feldman et al. [2015] as

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 \mid S = 0)}{\mathbb{P}(g(X) = 1 \mid S = 1)}. \tag{4.2.2}$$

The ideal scenario where g achieves statistical parity is equivalent to $DI(g, X, S) = 1$. As we have metioned, statistical parity is often unrealistic and we can consider instead a certain level of fairness as in the following definition.

Definition 4.2.1 *The classifier g has disparate impact at level $\tau \in (0,1]$, with respect to (X, S), if $DI(g, X, S) \leq \tau$.*

The disparate impact of a classifier measures its level of fairness: the smaller the value of τ, the less fair it is. In the following, we denote $a(g) := \mathbb{P}(g(X) = 1 \mid S = 0)$ and $b(g) := \mathbb{P}(g(X) = 1 \mid S = 1)$. In this paper, we will consider classifiers g such that $a(g) > 0$ and $b(g) > 0$ (the classifier is not totally unfair, in the sense that it does not predict the same outcome for a whole level of the protected attribute). Moreover, we assume $b(g) \geq a(g)$ (the default class $S = 1$ is more likely to have a successful outcome). Thus, in the definition above $0 < \tau \leq 1$. We point out that the value $\tau_0 = 0.8 = 4/5$, also known in the literature as the 80% *rule*, has been cited as a legal score to decide whether the discrimination of the algorithm is acceptable or not (see for instance Feldman et al. [2015]). This rule ensures that "for every 5 individuals with successful outcome in the majority class, 4 in the minority class will have a successful outcome too". It will be useful in the sequel to use the definition in the reverse (positive) sense: a classifier does not have disparate impact at level τ, with respect to (X, S), if $DI(g, X, S) > \tau$.

Finally, another definition has been proposed in the statistical literature on fair learning. Given a classifier $g \in \mathcal{G}$, its *balanced error rate* (BER) with respect to the joint distribution of the random vector (X, S) is defined as the average class-conditional error

$$BER(g, X, S) = \frac{a(g) + 1 - b(g)}{2}. \tag{4.2.3}$$

Notice that $BER(g, X, S)$ is the misclassification error of $g \in \mathcal{G}$ for predicting S when the protected classes are equally likely ($\mathbb{P}(S = 0) = \mathbb{P}(S = 1) = 1/2$). This allows to define the notion of $\varepsilon-$predictability of the protected attribute. S is said to be $\varepsilon-$predictable from X if there exists a classifier $g \in \mathcal{G}$ such that $BER(g, X, S) \leq \varepsilon$. Equivalently, S is not $\varepsilon-$predictable from X if $BER(g, X, S) > \varepsilon$, for all classifiers g chosen in the class $\mathcal{G}$. Thus, if $\min_{g \in \mathcal{G}} BER(g, X, S) = \varepsilon^*$ then S is not $\varepsilon-$predictable from X for all $\varepsilon < \varepsilon^*$.

In the following, we recast previous notions of fairness and provide a probabilistic framework to highlight the relationships between the distribution of the observations and the fairness of the classification problem. We denote $\mu_s := \mathcal{L}(X|S = s), s = 0, 1$. The following theorem generalizes the result in Feldman et al. [2015] showing the relationship between predictability, disparate impact and total variation distance.

Theorem 4.2.1 *Given r.v.'s $X \in \mathbb{R}^d$, $S \in \{0, 1\}$, the classifier g has disparate impact at level $\tau \in [0, 1]$, if and only if $BER(g, X, S) \leq \frac{1}{2} - \frac{a(g)}{2}(\frac{1}{\tau} - 1)$. Moreover*

$$\min_{g \in \mathcal{G}} BER(g, X, S) = \frac{1}{2}\left(1 - d_{TV}\left(\mu_0, \mu_1\right)\right).$$

As noted in the Introduction, to get rid of the possible discrimination associated to a classifier we could, in principle, either modify the classifier or the input data. If action on the algorithm is not possible (for instance, if we have no access to the values Y of the learning sample) we have to focus on the second option and change the data X to ensure that every classifier trained from the modified data would be fair with respect to S. This transformation aimed at breaking the dependence on the protected attribute, is called *repairing the data*. For this, Feldman et al. [2015], Johndrow and Lum [2019] or Hacker and Wiedemann [2017] propose to map the conditional distributions to a common distribution in order to achieve statistical parity. This *total repair* of the data amounts to modifying the input variables X building a repaired version, $\tilde{X}$, such that any classifier g trained from $\tilde{X}$ will have disparate impact $\tau = 1$, with respect to $(\tilde{X}, S)$ (equivalently, every classifier g that predicts Y from the new variable $\tilde{X}$ will achieve statistical parity). As a counterpart, it is clear that the choice of the target distribution should convey as much information as possible on the original variables, otherwise it would hamper the accuracy of the new classification.

In more detail, *total repair* amounts to mapping the original variable X into a new variable $\tilde{X} = T_S(X)$ such that conditional distributions with respect to S are the same, namely,

$$\mathcal{L}\left(\tilde{X} \mid S = 0\right) = \mathcal{L}\left(\tilde{X} \mid S = 1\right). \tag{4.2.4}$$

In this case, any classifier g built with such information will be such that $\mathcal{L}\left(g(\tilde{X}) \mid S = 0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S = 1\right)$, guaranteeing full fairness of the classification rule. To accomplish this transformation, the solution detailed in many papers is to map both conditional distributions μ_0 and μ_1 onto a common distribution ν. Actually, the distribution of X is modified using a random map $T_S : \mathbb{R}^d \to \mathbb{R}^d$ that depends on the value of the protected variable S and such that $\mathcal{L}(T_0(X) \mid S = 0) = \mathcal{L}(T_1(X) \mid S = 1)$. Consequently, two different problems arise.

- First of all, the choice of the distribution ν should be as similar as possible to both distributions μ_0 and μ_1 at the same time, in order to reduce the amount of information lost with this transformation, and thus still enabling the prediction task using the modified variable $\tilde{X} \sim \nu$ instead of the original X.

- Moreover, once the target ν is selected, we have to find the optimal way of transporting μ_0 and μ_1 into it.

First, from Theorem 4.2.1, the total variation distance is the natural choice to measure the distances between the conditional distributions in the fairness problem. However, this distance is computationally difficult to handle. Hence, previous works suggest the use of the Wasserstein metric, $\mathcal{W}_2$, which appears as an appropriate tool for comparing probability distributions and arises naturally in optimal transport theory. We refer to Villani [2009] for general background on the topic. In this framework, T_S will be a random transport map between the distributions $\mathcal{L}(X \mid S)$ and $\mathcal{L}(\tilde{X})$. Then, when considering an optimal choice for the target distribution for $\mathcal{L}(\tilde{X})$, some authors (see Feldman et al. [2015]) propose, in the one-dimensional case, to choose the distribution whose quantile is the mean of the quantile functions. In general this corresponds actually to the so-called Wasserstein barycenter of the laws $\mathcal{L}(X \mid S = s)$, as we describe next.

Given probability measures $(\mu_j)_{1 \leq j \leq J}$ with finite second moment and weights $(\omega_j)_{1 \leq j \leq J}$, the Wasserstein barycenter is a minimizer of

$$\nu \mapsto \sum_{j=1}^{J} \omega_j \mathcal{W}_2^2(\nu, \mu_j), \tag{4.2.5}$$

see Agueh and Carlier [2011]. Empirical versions of the barycenter and their properties are analyzed in Boissard et al. [2015] or Le Gouic and Loubes [2017]. Similar ideas have also been developed in Cuturi and Doucet [2014] or del Barrio and Loubes [2019]. In general, the Wasserstein barycenter appears to be a meaningful feature to represent the mean prototype of a set of distributions. Note that in the one dimensional case, the mean of the quantile functions corresponds actually to the minimizer of (4.2.5).

In the following section, we present some statistical justifications for this choice. Computation of Wasserstein barycenters may be a difficult issue in the general case. Yet, in this work we only consider the barycenter between two probabilities μ_0, μ_1 on $\mathbb{R}^d$, so we provide some details on how to compute this barycenter in general dimension.

4.3 Repair with Wasserstein Barycenter

4.3.1 Learning with Wasserstein Barycenter distribution

In our particular problem, where $J = 2$ in (4.2.5), the conditional distributions μ_0 and μ_1 are going to be transformed into the distribution of the Wasserstein barycenter μ_B between them, with weights π_0 and π_1, defined as

$$\mu_B \in argmin_{\nu \in \mathcal{P}_2} \left\{ \pi_0 \mathcal{W}_2^2(\mu_0, \nu) + \pi_1 \mathcal{W}_2^2(\mu_1, \nu) \right\}.$$

Let $\tilde{X}$ be the transformed variable with distribution μ_B. For each $s \in \{0, 1\}$, the deformation will be performed through the optimal transport map (o.t.m.) $T_s : \mathbb{R}^d \to \mathbb{R}^d$ pushing each μ_s towards the weighted barycenter μ_B. The existence of μ_B is guaranteed (see Theorem 2.12 in Villani [2003]) as soon as μ_s are absolutely continuous (a.c.) with respect to Lebesgue measure. In that case,

$$\mathbb{E}\left(\|X - T_s(X)\|^2 \mid S = s \right) = \mathcal{W}_2^2(\mu_s, \mu_B). \tag{4.3.1}$$

Remark 4.3.1 *Note that computing the barycenter of two measures is equivalent to the computation of the o.t.m. between them. If μ_0 is a.c. on $\mathbb{R}^d$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ denotes the o.t.m. between μ_0 and μ_1, that is $\mu_1 = \mu_{0\sharp}T$, then $\mu_\lambda = \mu_{0\sharp}\left((1-\lambda)Id + \lambda T\right)$ is the weighted barycenter between μ_0 and μ_1, with weights $1-\lambda$ and λ, respectively. The map $(1-\lambda)Id+\lambda T$ is an optimal transport plan for all $\lambda \in [0,1]$. So, the complexity of computing $\mu_B = \mu_{0\sharp}\left(\pi_0 Id + \pi_1 T\right)$ is the same as computing T.*

Remark 4.3.2 *Note also that for distributions on the real line, we can write the explicit expression of the barycenter μ_B based on the exact solution to the optimization problem (4.3.1). Given $s \in \{0,1\}$ and $X \in \mathbb{R}$, let $F_s : \mathbb{R} \to [0,1]$ denote the cumulative distribution function of X, given $S = s$, and $F_s^{-1} : [0,1] \to \mathbb{R}$ its quantile associated function. The weighted Wasserstein barycenter μ_B of μ_0 and μ_1 is the unique minimizer of the functional (4.2.5) and its quantile function can be computed as*

$$F_B^{-1}(t) = \left(\lambda F_0^{-1}(t) + (1-\lambda)F_1^{-1}(t)\right), \ t \in [0,1].$$

Moreover, note that $F_s\left(X \mid S = s\right) \sim \mathcal{U}(0,1), s = 0,1$, and the o.t.m. solution to (4.3.1) is $T_s = F_B^{-1} \circ F_s$.

To understand the use of the Wasserstein barycenter as the target distribution for μ_0 and μ_1, we will quantify the amount of information lost when replacing the distribution of X by a new and, for the moment, unknown distribution of $\tilde{X}$ obtained by transporting μ_0 and μ_1. Set the random transport plan $T_S : \mathbb{R}^d \to \mathbb{R}^d$, and the modified variable $\tilde{X} = T_S(X)$. We point out that choosing the distribution of $\tilde{X}$ amounts to choosing the transport plans T_0 and T_1. We are facing learning problems in two different settings.

- On the one hand, the full information available are the input variables X and the protected variable S, which play an important role in the classification, since the classifier has a different behavior according to the different classes $S = 0$ and $S = 1$. Hence, we let S play a role in the decision process since it is associated to Y, and possibly giving rise to a different treatment for the two different groups. In this case, the classification risk when the full data (X, S) is available can be computed as $R(g, X, S)$, the risk in the prediction of a classification rule g that depends on both variables X and S, namely $R(g, X, S) := \mathbb{P}(g(X, S) \neq Y)$
- On the other hand, in the repair data only the modified version $\tilde{X}$ of the input is at hand. Hence, the risk when learning a classifier is $R(h, \tilde{X}) := \mathbb{P}(h(\tilde{X}) \neq Y)$.

Studying the efficiency of the method requires providing a bound for the difference between the minimal risks obtained for the best classifier with input data $\tilde{X} = T_S(X)$, and for the best classifier with input data (X, S), called g_B. These risks are respectively denoted $R_B(\tilde{X})$ and $R_B(X, S) = \inf_g R(g, X, S) = R(g_B, X, S)$, and then its difference is

$$\mathcal{E}(\tilde{X}) := R_B(\tilde{X}) - R_B(X, S).$$

Note first that, given $X = x$ and $S = s$, $\inf_g R(g, X, S)$ can be computed by mimicking the usual expression of the 2-class classification error as in Bousquet et al. [2004], for instance. Denoting by $\eta_s(x) := \mathbb{P}(Y = 1 \mid X = x, S = s)$,

$$\begin{aligned}&\mathbb{P}(g(X, S) \neq Y \mid X = x, S = s)\\ &= \mathbb{1}_{g(x,s)\neq 0}(1 - \eta_s(x)) + \mathbb{1}_{g(x,s)\neq 1}\eta_s(x).\end{aligned}$$

So we deduce that $R(g,X,S)=\mathbb{E}\left[\mathbb{1}_{g(X,S)=0}(2\eta_S(X)-1)\right]+\mathbb{E}\left[1-\eta_S(X)\right]$. The minimum risk is thus obtained using the Bayes' rule $g_B(x,s)=\mathbb{1}_{\eta_s(x)>1/2}$, showing that

$$\begin{aligned}R_B(X,S) &:= min_g R(g,X,S)\\ &= \mathbb{E}\left[\mathbb{1}_{\{2\eta_S(X)-1<0\}}(2\eta_S(X)-1)\right]+\mathbb{E}\left[1-\eta_S(X)\right].\end{aligned}$$

Similarly, the risk related to a classifier $h(\tilde{X})$ is given by

$$\begin{aligned}R(h,\tilde{X}) &= R(h,T_S(X))\\ &= \mathbb{E}\left[\mathbb{1}_{h\circ T_S(X)=0}(2\eta_S(X)-1)\right]+\mathbb{E}\left[1-\eta_S(X)\right].\end{aligned} \tag{4.3.2}$$

Hence, the amount of information lost due to modifying the data is controlled by the following theorem.

Theorem 4.3.3 *Consider $X\in\mathbb{R}^d$ and $S\in\{0,1\}$. Let $T_S:\mathbb{R}^d\to\mathbb{R}^d$, $d\geq 1$ be a random transformation such that $\mathcal{L}(T_0(X)\mid S=0)=\mathcal{L}(T_1(X)\mid S=1)$, and consider $\tilde{X}=T_S(X)$. Assume that $\eta_s(X)$ is Lipschitz with constant $K_s>0$, $s=0,1$. Then, if $K=\max\{K_0,K_1\}$,*

$$\mathcal{E}(\tilde{X})\leq 2\sqrt{2}K\left(\sum_{s=0,1}\pi_s\mathcal{W}_2^2(\mu_s,\mu_{s\sharp}T_s)\right)^{\frac{1}{2}}. \tag{4.3.3}$$

Theorem 4.3.3 provides some justification to the use of the Wasserstein barycenter as the distribution of the modified variable. Similar inequalities in the framework of domain adaptation are given in Redko et al. [2017]. In fact, minimizing the upper bound in (4.3.3) with respect to the function $T_S:\mathbb{R}^d\to\mathbb{R}^d$, $d\geq 1$, leads to consider the transport plan carrying the conditional distributions μ_0 and μ_1 towards their Wasserstein barycenter μ_B with weights π_0,π_1, that is, $\mu_{S\sharp}T_S=\mu_B$. Hence, this provides some understanding on the choice of the Wasserstein barycenter advocated in the work Feldman et al. [2015] and leads to the following bound

$$\begin{aligned}&\inf_{T_S}\left\{R(g_B\circ T_S,X)-R(g_B,X,S)\right\}\\ &\leq 2\sqrt{2}K\left(\sum_{s=0,1}\pi_s\mathcal{W}_2^2(\mu_s,\mu_B)\right)^{\frac{1}{2}}\leq\frac{K}{\sqrt{2}}\mathcal{W}_2^2(\mu_0,\mu_1).\end{aligned}$$

This upper bound only provides some guidelines on the choice of the target distribution. Nevertheless, choosing the Wasserstein barycenter provides a reasonable and, more important, feasible solution to achieve fairness. Recently in del Barrio et al. [2019b] a CLT for L_p transportation cost in $\mathbb{R}$ is provided, which enables to build two sample tests and confidence intervals to certify the similarity between two distributions. We also point out that we only deal with the case of 2 classes for S, a majority and a minority, which is one of the main concerns in fair learning. Yet, the result could be generalized to multiclass where $S\in\mathcal{S}$ with several labels since it only relies on the defintion of the Wasserstein barycenter. In this case, computing the barycenter becomes a harder issue.

As pointed out previously, the *total repair* process ensures full fairness but at the expense of the accuracy of the classification. A solution for this could be found in Feldman et al. [2015], called *geometric repair*. The authors propose not to move the conditional distributions to the barycenter but only partly towards it along the Wasserstein's geodesic path between μ_0 and μ_1. We analyze next this procedure and propose an alternative method for the partial repair.

4.3.2 A new algorithm for partial repair

Let $\lambda \in [0,1]$ be the parameter representing the amount of repair desired for X. Let Z be a target variable with distribution μ. Set $R_s = T_s^{-1},\ s = 0,1$, where T_s is the o.t.m. pushing each μ_s towards the target μ. Note that $R_s(Z)$ follows the original conditional distribution μ_s.

Definition 4.3.1 (Random repair) *Let B be a Bernoulli variable with parameter λ. With the above notation, we define for $s \in \{0,1\}$, and $\lambda \in (0,1)$ the repaired distributions*

$$\begin{aligned} \tilde{\mu}_{s,\lambda} &= \mathcal{L}(BZ + (1-B)R_s(Z)) \\ &= \mathcal{L}(BT_s(X) + (1-B)X \mid S = s). \end{aligned} \tag{4.3.4}$$

This repair procedure consists in randomly changing the distribution of the original X by either selecting the target μ or the original conditional distributions. The degree of repair is governed by the Bernoulli parameter λ: note that for $\lambda = 0$ $\tilde{\mu}_{s,0} = \mathcal{L}(X \mid S = s)$ and for $\lambda = 1$ $\tilde{\mu}_{s,1} = \mathcal{L}(Z) = \mu$. The value of λ should come from a trade-off between *(i)* the accuracy of the new classification result, that leads to little changes in the initial distributions; and *(ii)* the non-predictability of the protected variable, which implies that the two conditional distributions should stay close with respect to the total variation distance. In fact, (see e.g. Massart [2007]), the distance in total variation between two probabilities P and Q can be computed as

$$d_{TV}(P,Q) = \min_{\pi \in \Pi(P,Q)} \pi(x \neq y). \tag{4.3.5}$$

This leads to

$$\begin{aligned} &d_{TV}(\tilde{\mu}_{0,\lambda}, \tilde{\mu}_{1,\lambda}) \leq \mathbb{P}(BZ + (1-B)R_0(Z) \\ &\neq BZ + (1-B)R_1(Z)) = 1 - \mathbb{P}(BZ + (1-B)R_0(Z) \\ &= BZ + (1-B)R_1(Z)) \leq 1 - \mathbb{P}(B = 1) = 1 - \lambda. \end{aligned}$$

This bound suggests that λ should be close to 1 to ensure non-predictability of S. Finally, observe that the misclassification error using the randomly repaired data is a mixture of the two errors with the totally repaired variable $T_S(X)$ and the original X since $R(g, \tilde{X}_\lambda) = (1-\lambda)\mathbb{P}(g(X) \neq Y) + \lambda\mathbb{P}(g(T_S(X)) \neq Y)$. Hence, from Theorem 4.3.3 the use of the Wasserstein barycenter $Z \sim \mu_B$ is justified.

In the literature (for instance Zafar et al. [2017a]), another partial repair procedure is used, called *geometric repair*. As before, μ is chosen as the barycenter μ_B and the partially repaired conditional distributions are defined as

$$\begin{aligned} \mu_{s,\lambda} &= \mathcal{L}(\lambda Z + (1-\lambda)R_s(Z)) \\ &= \mathcal{L}(\lambda T_s(X) + (1-\lambda)X \mid S = s),\ s \in \{0,1\}. \end{aligned}$$

Observe that $\lambda = 1$ yields the fully repaired variable, and $\lambda = 0$ leaves the conditional distributions unchanged. So the parameter λ governs how close the distributions are to the barycenter. Such procedure sounds appealing since the conditional distributions are moved on the geodesic path between the original distributions which warrants an optimal prediction and the barycenter which warrants fairness. Controling this distance is the key of the *geometric repair*. Yet, reasoning among the lines of previous argument to obtain an upper bound for the classification risk using the partially repaired distributions $\mu_{0,\lambda}$ and $\mu_{1,\lambda}$ does not lead to a satisfying result. This comes from the fact that the *geometric repair* moves the original distributions according

to the Wasserstein distance, while fairness is measured through the total variation distance, and they are of different nature. So if $\lambda \in (0,1)$, using (4.3.5) implies that

$$\begin{aligned} d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) &\leq \mathbb{P}(\lambda Z + (1-\lambda)R_0(Z) \\ &\neq \lambda Z + (1-\lambda)R_1(Z)) = \mathbb{P}(R_0(Z) \neq R_1(Z)). \end{aligned} \tag{4.3.6}$$

The previous bound means that the amount of repair quantified by λ does not affect the TV distance between the modified conditional distributions. Moreover, in some situations, (4.3.6) turns out to be an equality. Consider, for instance,

$$\mu_{0,0} = U(K, K+1) \qquad \mu_{1,0} = U(-K-1, -K) \tag{4.3.7}$$

as the distributions of X in each class. Then, the barycenter is $\mu_{0,1} = \mu_{1,1} = U(-1/2, 1/2)$ and

$$\begin{aligned} \mu_{0,\lambda} &= U\left(-\frac{\lambda}{2} + (1-\lambda)K, -\frac{\lambda}{2} + (1-\lambda)K + 1\right), \\ \mu_{1,\lambda} &= U\left(-\frac{\lambda}{2} - (1-\lambda)(K+1), -\frac{\lambda}{2} - (1-\lambda)(K+1) + 1\right). \end{aligned}$$

In this case, the TV distance can be easily computed as

$$d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) = \min(1, (1-\lambda)(2K+1)). \tag{4.3.8}$$

We see from equation (4.3.8) that $d_{TV}(\mu_{0,\lambda}, \mu_{1,\lambda}) = 1$, if $\lambda \leq 2K/(2K+1)$, which means that the protected attribute could be perfectly predicted from the partially repaired data set for values of λ arbitrarily close to 1. Thus, the bound (4.3.6) provides some argument against the *geometric* method since the repair should favour small distances between the original distributions to ensure a certain desired level of fairness. Hence, rather than using a displacement along the Wasserstein geodesic between the distributions, we propose the *random repair* which enables a better control of their total variation distance, enhancing the disparate impact while not hampering too much the efficiency of the classification.

In the next section, we propose a new algorithm for the *total repair* which in practice attains full fairness in contrast with the existing in the literature. Based on it, we design a scheme to perform the *random repair*.

4.4 Computational aspects for Repairing Datasets in General Dimension

Let $\{(X_i, S_i, Y_i), i = 1, \ldots, N\}$ be an observed sample of (X, S, Y), and denote by n_0 and n_1 the number of instances in each protected class. Without loss of generality, we assume that the observations are ordered by the value of S,

$$\begin{aligned} &x_{0,i} := X_i, \text{ if } s_i = 0, \ i = 1, \ldots, n_0, \\ &x_{1,j-n_0} := X_j, \text{ if } s_j = 1, \ j = n_0+1, \ldots, N = n_0 + n_1. \end{aligned}$$

Generally, the sizes of the samples $\mathcal{X}_0 = \{x_{0,1}, \ldots, x_{0,n_0}\}$ and $\mathcal{X}_1 = \{x_{1,1}, \ldots, x_{1,n_1}\}$ are different and Monge maps may not even exist between an empirical measure to another. This happens when their weight vectors are not compatible, which is always the case when the target measure has more points than the source measure. Hence, the solution to the optimal transport problem does not correspond to finding an optimal transport map, but an optimal transport distribution.

The cuadratic cost function becomes discrete as it can be written as a matrix $C = (c_{ij})$, with $c_{ij} = \|x_{0,i} - x_{1,j}\|^2, 1 \leq i \leq n_0,\ 1 \leq j \leq n_1$. When $\mu_{0,n} = \sum_{i=1}^{n_0} \frac{1}{n_0}\delta_{x_{0,i}}$ and $\mu_{1,n} = \sum_{j=1}^{n_1} \frac{1}{n_1}\delta_{x_{1,j}}$, the Wasserstein distance $\mathcal{W}_2(\mu_{0,n}, \mu_{1,n})$ between them is the squared root of the optimum of a net-work flow problem known as the *transportation problem*. It consists in finding a matrix $\gamma \in \mathcal{M}_{n_0 \times n_1}(\mathbb{R})$ which minimizes the transportation cost between the two distributions as follows

$$\begin{cases} \min_\gamma \sum\limits_{\substack{1\leq i\leq n_0 \\ 1\leq j\leq n_1}} c_{ij}\gamma_{ij}, & \text{subject to:} \\ \gamma_{ij} \geq 0, & \\ \sum_{i=1}^{n_0} \gamma_{ij} = \frac{1}{n_1}, & \text{for all } j, \\ \sum_{j=1}^{n_1} \gamma_{ij} = \frac{1}{n_0}, & \text{for all } i. \end{cases} \tag{4.4.1}$$

If $\hat{\gamma}$ is a solution to the linear program (4.4.1) then, the measure

$$\mu_{B,n} = \sum_{\substack{1\leq i\leq n_0 \\ 1\leq j\leq n_1}} \hat{\gamma}_{ij}\delta_{\{\pi_0 x_{0,i} + \pi_1 x_{1,j}\}}$$

is a barycenter of $\mu_{0,n}$ and $\mu_{1,n}$, with weights π_0 and π_1, according to Remark 4.3.1. See Cuturi and Doucet [2014] for details on the discrete Wasserstein and Optimal Transport computation.

4.4.1 Total repair

In practice, the implementation of the repair scheme in Section 4.3 is based on the transport matrix $\hat{\gamma}$ from $\mathcal{X}_0$ to $\mathcal{X}_1$. As we have pointed out, in this transport scheme the major difficulty comes from the fact that the sizes of these sets are different and the transport is not a one-by-one mapping. Each point in the source set could be transported (with weights) into several points of the target, or various points in the source could be moved into the same point of the target. As a consequence, we must adapt the algorithm that produces the repaired data set, denoted by $\tilde{\mathcal{X}}$.

We detail next two different methods. The first one is similar to some existing in the literature and does not achieve total fairness in practice, while the second one is a novelty and does guarantee this property for the new data $\tilde{\mathcal{X}}$.

(A) As depicted in Figure 4.3(A), each original point in $\mathcal{X}_0, \mathcal{X}_1$ is changed by a unique point given by

$$\tilde{x}_{0,i} = \pi_0 x_{0,i} + n_0\pi_1 \sum_{j=1}^{n_1} \gamma_{ij}x_{1,j},\ 1 \leq i \leq n_0,$$
$$\tilde{x}_{1,j} = n_1\pi_0 \sum_{i=1}^{n_0} \gamma_{ij}x_{0,i} + \pi_1 x_{1,j},\ 1 \leq j \leq n_1.$$

The set $\tilde{\mathcal{X}}$ will be a collection of exactly $n_0 + n_1$ points. This approach generalizes to higher dimensions the idea in Feldman et al. [2015] and Johndrow and Lum [2019], which only considered the unidimensional case, where the transport is written in terms of the distribution funtions. Yet, in practice it builds two different sets $\tilde{\mathcal{X}}_0 = \{x_{0,i},\ 1 \leq i \leq n_0\}$ and $\tilde{\mathcal{X}}_1 = \{x_{1,j},\ 1 \leq j \leq n_1\}$ that do not ensure (4.2.4).

(B) To ensure total fairness, each point will split its mass to be transported into several modified versions. This generates an extended set $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_0 \cup \tilde{\mathcal{X}}_1$, which is formed by the complete

distribution $\mu_{B,n}$. As shown in Figure 4.3(B), if $\hat{\gamma}_{ij} > 0, 1 \leq i \leq n_0,\ 1 \leq j \leq n_1$, we define the two points

$$\tilde{x}_{0,i,j} := \tilde{x}_{1,j,i} = \pi_0 x_{0,i} + \pi_1 x_{1,j}, \tag{4.4.2}$$

and the sets

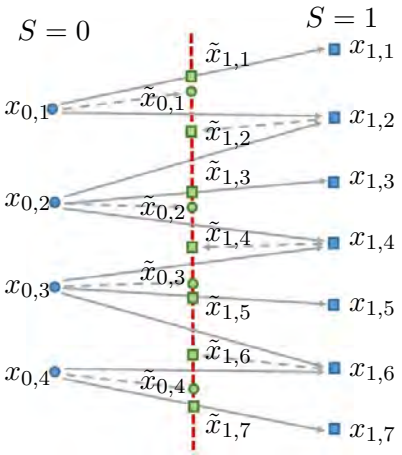
$$\tilde{\mathcal{X}_0} := \bigcup_{1 \leq i \leq n_0} \{\tilde{x}_{0,i,j} \ / \ \hat{\gamma}_{ij} > 0, 1 \leq j \leq n_1\}$$
$$\tilde{\mathcal{X}_1} := \bigcup_{1 \leq j \leq n_1} \{\tilde{x}_{1,j,i} \ / \ \hat{\gamma}_{ij} > 0, 1 \leq i \leq n_0\}.$$

The rebuilt distributions have sizes equal to the number of non zero elements in $\hat{\gamma}$, and each point has weight $\hat{\gamma}_{ij}$. Unlike the previous, this approach does achieve total impredictability, as it manages to produce repaired conditional distributions equally distributed.

Example 4.4.1 *We have simulated two samples $\mathcal{X}_0$ and $\mathcal{X}_1$ of points in $\mathbb{R}$ of sizes $n_0 = 4$ and $n_1 = 7$. The optimal matrix solution to the problem* (4.4.1) *is*

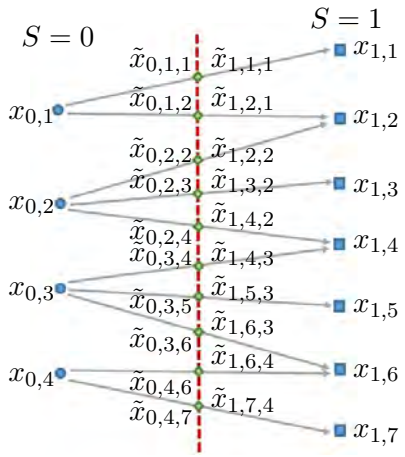
$$\hat{\gamma} = \begin{bmatrix} \frac{1}{7} & \frac{1}{4} - \frac{1}{7} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2}{7} - \frac{1}{4} & \frac{1}{7} & \frac{1}{14} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{14} & \frac{1}{7} & \frac{2}{7} - \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} - \frac{1}{7} & \frac{1}{7} \end{bmatrix}$$

*If $\mathcal{X}_0$ and $\mathcal{X}_1$ are realizations of μ_0 and μ_1, respectively, then the left part of Figure 4.3 represents procedure **(A)** that produces the repaired sets $\tilde{\mathcal{X}_0} = \{\tilde{x}_{0,1}, \ldots, \tilde{x}_{0,4}\}$ (rounded green points) and $\tilde{\mathcal{X}_1} = \{\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,7}\}$ (squared green points). As we can observe, the two sets are clearly different and the statistical parity can not be reached. Otherwise, procedure **(B)** on the right yields to $\tilde{\mathcal{X}_0} = \tilde{\mathcal{X}_1}$.*



(A)

Figure 4.1



(B)

Figure 4.2

Remark 4.4.1 *When the two samples $\mathcal{X}_0$ and $\mathcal{X}_1$ have equal size n and the weights $\gamma_{ij} = \frac{1}{n}$, $1 \leq i, j \leq n$, are uniform, the mass conservation constraint implies that γ is a bijection and the Monge problem is equivalent to the optimal matching problem $\min_{\sigma \in Perm(n)} \frac{1}{n} \sum_{i=1}^{n} c_{i,\sigma(i)}$. Both repairing procedures **(A)** and **(B)** perfom the same generating $\tilde{x}_{0,i} = \tilde{x}_{1,i} = \frac{1}{2}(x_{0,i} + x_{1,i})$, $1 \leq i \leq n$, as depicted in Figure 4.4. Then, total fairness is always achieved.*

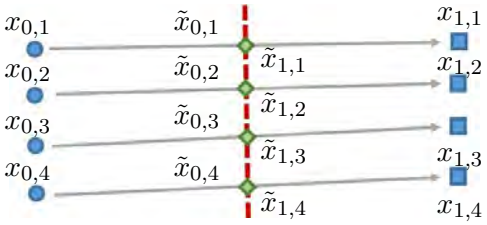


Figure 4.4 – Repairing process when $n_0 = n_1$

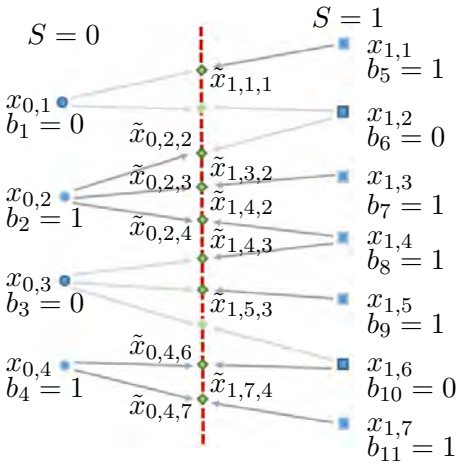


Figure 4.5 – Example of the *random repair* with $\lambda = \frac{1}{2}$.

4.4.2 Random repair

As previously noted, trying to build the set $\tilde{\mathcal{X}}$ satisfying the goal (4.2.4) may compromise too much the accuracy of the classification with these new data. In this sense, the *random repair* procedure proposed in this paper aims at setting a tradeoff between fairness and accuracy through the parameter λ, that models the amount of repair desired. We detail next how to compute the randomly repaired set denoted by $\tilde{\mathcal{X}}_\lambda$, with $\lambda \in [0, 1]$. According to (4.3.4), we will randomly select either the points in the original sets $\mathcal{X}_0$ and $\mathcal{X}_1$ or their repaired sequels with procedure **(B)**. For this, consider a sample $b_1, \ldots, b_{n_0+n_1} \sim B(\lambda)$, and define

$$\tilde{\mathcal{X}}_{0,\lambda} := \bigcup_{i=1}^{n_0} R_{0,i,\lambda} \qquad \tilde{\mathcal{X}}_{1,\lambda} := \bigcup_{j=1}^{n_1} R_{1,j,\lambda}, \tag{4.4.3}$$

where $R_{0,i,\lambda}$ and $R_{1,j,\lambda}$ are the repaired sets of the points $x_{0,i}$ and $x_{1,j}$, respectively:

$$R_{0,i,\lambda} := \begin{cases} \{x_{0,i}\} & \text{if } b_i = 0 \\ \{\tilde{x}_{0,i,j} \ / \ \hat{\gamma}_{ij} > 0, 1 \le j \le n_1\} & \text{if } b_i = 1 \end{cases}$$
$$R_{1,j,\lambda} := \begin{cases} \{x_{1,j}\} & \text{if } b_{n_0+j} = 0 \\ \{\tilde{x}_{1,j,i} \ / \ \hat{\gamma}_{ij} > 0, 1 \le i \le n_0\} & \text{if } b_{n_0+j} = 1 \end{cases}$$

with $\tilde{x}_{0,i,j}$ and $\tilde{x}_{1,j,i}$ given in (4.4.2), with weights $\hat{\gamma}_{i,j}$.

Example 4.4.2 *Consider the situation in Example 4.4.1. Figure 4.5 represents the random repair procedure for $\lambda = \frac{1}{2}$. For $l = 1, \ldots, n_0 + n_1 = 11$, we have simulated values $b_l \sim \mathcal{B}(\frac{1}{2})$. From* (4.4.3) *we have the randomly repaired sets*

$$\tilde{\mathcal{X}}_{0,\lambda} = \{x_{0,1}, \tilde{x}_{0,2,2}, \tilde{x}_{0,2,3}, \tilde{x}_{0,2,4}, x_{0,3}, \tilde{x}_{0,4,6}, \tilde{x}_{0,4,7}\}$$
$$\tilde{\mathcal{X}}_{1,\lambda} = \{\tilde{x}_{1,1,1}, x_{1,2}, \tilde{x}_{1,3,2}, \tilde{x}_{1,4,2}, \tilde{x}_{1,4,3}, \tilde{x}_{1,5,3}, x_{1,6}, \tilde{x}_{1,7,4}\}.$$

4.5 Application with simulated data

In this section, we present an application of the repairing procedures in Section 4.3 to some simulated data to illustrate their performance. We also provide an example in which the *geometric repair* fails to remove the bias in the data.

To introduce some bias in the simulated dataset $\mathcal{X}$ we have taken $n_0 = 600$ and $n_1 = 400$ examples from two multivariate normal distributions on $\mathbb{R}^5$ with vector of means $\mu_0 = (3, 3, 2, 2.5, 3.5)$ and $\mu_1 = (4, 4, 3, 3.5, 4.5)$ and equal covariance matrices $\Sigma = diag(1, 1, 0.5, 0.5, 1)$. Then, in order to simulate the classification Y, we have chosen parameters $\beta_0 = (1, -1, -0.5, 1, -1, 1)$ and $\beta_1 = (1, -0.4, 1, -1, 1, -0.5)$ to build a logit model for each group with different probability of success for $s = 0, 1$, $\pi_s(x) = \frac{e^{X\beta_s}}{1+e^{X\beta_s}}$, higher for the class $S = 1$.

Then, a new logit classifier has been trained from this simulated data, splitting the set into the learning and the test sample using the ratio 300 / 700. In the first row of Table 4.3 we can see a summary of the performance of the logit with the original data. We have estimated the disparate impact using its empirical counterpart and provided a confidence interval which was established in Besse et al. [2018b]. Before the repair, we can say with a confidence of 95% that the logit rule has DI at level 0.53 with respect to S. Then, we have made the repair in $\mathbb{R}^5$ in the testing sample using the different procedures studied in this paper. We have used the previous logit model, which was trained from biased data, to classify such repaired observations. In the remaining rows of Table 4.3 a summary of the performance of the logit with the repaired data using procedures **(A)** and **(B)** is presented. We note that in the experiments with procedure **(A)** the estimated value for DI is not exactly 1, as we have already anticipated. On the other hand, procedure **(B)** manages to change the data to attain statistical parity. The error in the logit classification done with the repaired data sets is a bit higher for the second procedure.

Finally, we present some results of the performance of the *Geometric* and *random repair*. Figure 4.6 represents the evolution of the confidence interval for the disparate impact with the amount of repair $0 \leq \lambda \leq 1$. Figure 5.5 shows the evolution with λ of the error in the classification done from the modified data set. For the experiments concerning the *random repair* procedure, we have repeated it 100 times and then we have computed the mean of the simulations. Clearly, the reached level of DI of the logit rule is higher with the *random repair*. We note that the amount of repair necessary to achieve an estimated DI at level 0.8 for the logit rule is 0.475 with the *random repair*, which entails an error of 0.1537; and 0.7 with the *geometric repair*, which entails an error of 0.1371.

Table 4.1 – Disparate impact of the logit with the original and the repaired datasets

Repair	Error	Difference	$\hat{DI}$	CI 95%
-	0.0943	-	0.5309	$(0.4230, 0.6389)$
(A)	0.1629	0.0686	0.9588	$(0.7641, 1.1535)$
(B)	0.1874	0.0931	1	$(0.8536, 1.1464)$

In order to see the failure of the *geometric repair*, we have simulated $n_0 = n_1 = 500$ observations from uniform distributions as in (4.3.7) with $K = 10$. We have trained a random forest classifier with the same ratio 300/700 for the learning and test sample. In Figure 4.8 we can see that the evolution of the disparate impact is controlled by the amount of repair only if we use the *random repair*. As pointed out from inequality (4.3.8), we observe that for values of $\lambda \leq \frac{20}{21} \approx 0.95$, the DI does not increase with λ for the partially modified distributions with the *geometric repair*. This means that for values of the degree of repair close to 1, this procedure does not manage to remove the bias in the data and consequently, it does not ensure the fairness of every classifier.

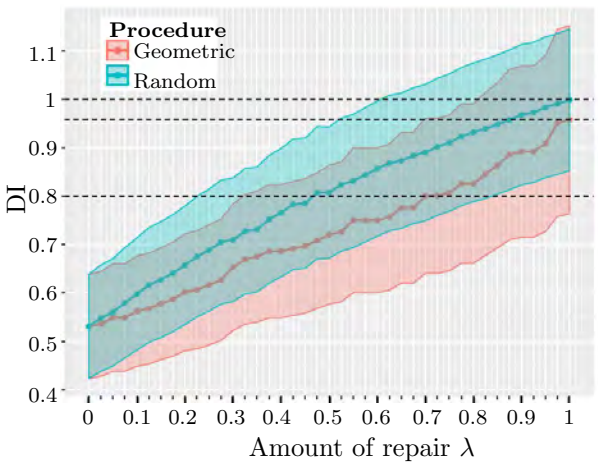


Figure 4.6 – CI at level 95% for DI of the logit

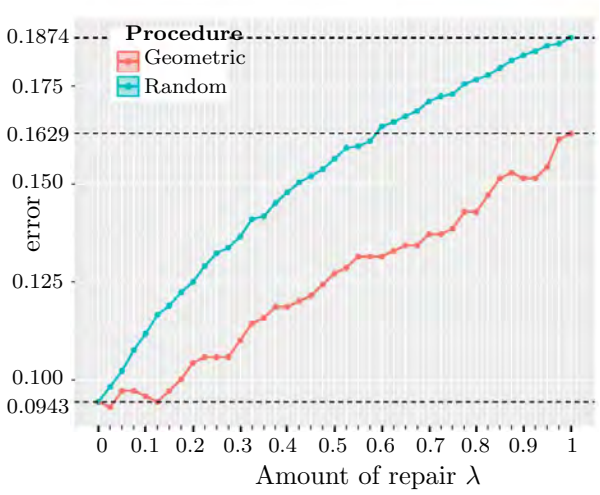


Figure 4.7 – Error of the logit

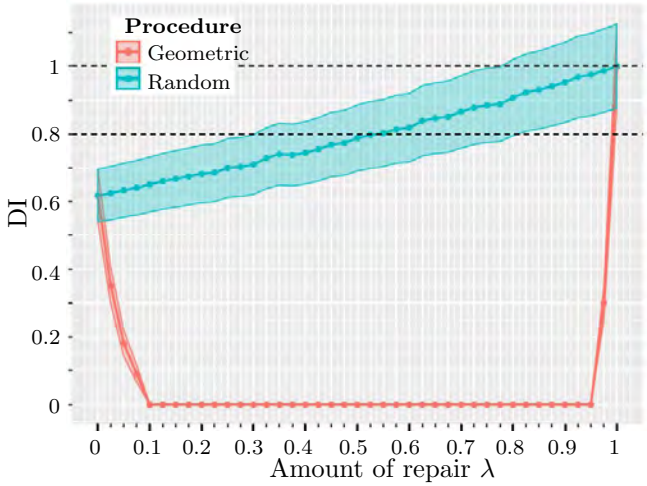


Figure 4.8 – CI at level 95% for DI of the random forest classifier

4.6 Conclusions

We have provided a multidimensional expansion and a feasible algorithm to repair a learning sample and incorporate fairness to prevent unfair algorithms to be learnt. Moreover this way of correction can be improved using a random reparation as shown in the paper. Yet this way of reparation only deals with disparate impact assessment and other criterion such as conditional accuracy equality for instance will be further incorporated using the same ideas of Wasserstein barycenter of conditional distributions.

4.7 Appendix A to Chapter 4

4.7.1 Proofs

Proof of Theorem 4.2.1. We will show that the conditions $DI(g,X,S) \leq \tau$ and $BER(g,X,S) \leq \frac{1}{2} - \frac{a(g)}{2}(\frac{1}{\tau} - 1)$ are equivalent, for all $g \in \mathcal{G}$. Indeed, given $g \in \mathcal{G}$,

$$
\begin{aligned}
BER(g,X,S) \leq \frac{1}{2} - \frac{a(g)}{2}\left(\frac{1}{\tau} - 1\right) = \frac{1}{2} - \frac{(\frac{1}{\tau}-1)}{2}\mathbb{P}(g(X) = 1 \mid S = 0) \\
\Leftrightarrow \mathbb{P}\left(g(X) = 0 \mid S = 1\right) + \mathbb{P}\left(g(X) = 1 \mid S = 0\right) \leq 1 - \left(\frac{1}{\tau} - 1\right)\mathbb{P}(g(X) = 1 \mid S = 0) \\
\Leftrightarrow \left(1 + \left(\frac{1}{\tau} - 1\right)\right)\mathbb{P}\left(g(X) = 1 \mid S = 0\right) + \mathbb{P}\left(g(X) = 0 \mid S = 1\right) \leq 1 \\
\Leftrightarrow \frac{1}{\tau}\mathbb{P}\left(g(X) = 1 \mid S = 0\right) \leq 1 - \mathbb{P}\left(g(X) = 0 \mid S = 1\right) = \mathbb{P}\left(g(X) = 1 \mid S = 1\right) \\
\Leftrightarrow DI(g,X,S) = \frac{\mathbb{P}\left(g(X) = 1 \mid S = 0\right)}{\mathbb{P}\left(g(X) = 1 \mid S = 1\right)} \leq \tau.
\end{aligned}
$$

Moreover, we denote by $f_i, i = 0, 1$, the density functions of the conditioned variables $X/S = i$, respectively, whose corresponding probability measures are both supposed to be, without loss of generality, absolute continuous with respect to a measure μ. In general, the misclassification error could be written as:

$$
\begin{aligned}
\mathbb{P}(g(X) \neq S) = \mathbb{P}(S = 0)\mathbb{P}\left(g(X) = 1 \mid S = 0\right) + \mathbb{P}(S = 1)\mathbb{P}\left(g(X) = 0 \mid S = 1\right) = \\
\mathbb{P}(S = 0)\int_{g(X)=1} f_0(x)d\mu(x) + \mathbb{P}(S = 1)\int_{g(X)=0} f_1(x)d\mu(x). \quad (4.7.1)
\end{aligned}
$$

Now, for $s = 0, 1$, we fixe the value of $\pi_s = \mathbb{P}(S = s)$, and from the Bayes' Formula, we know that

$$\mathbb{P}\left(S = s|X\right) = \frac{\pi_s f_s(X)}{\pi_0 f_0(X) + \pi_1 f_1(X)}.$$

Hence,

$$\left\{\mathbb{P}\left(S = 0|X\right) > \mathbb{P}\left(S = 1|X\right)\right\} = \left\{\pi_0 f_0(X) > \pi_1 f_1(X)\right\},\ \mu - a.s.$$

Thus, we can deduce that the classifier that minimizes the missclassification error rate is

$$g^*(x) = \begin{cases} 1 & if \quad \pi_0 f_0(x) \leq \pi_1 f_1(x) \\ 0 & if \quad \pi_0 f_0(x) > \pi_1 f_1(x) \end{cases},$$

and from equation (4.7.1),

$$\min_{g\in\mathcal{G}} \mathbb{P}(g(X) \neq S) = \int_{\{\pi_0 f_0(x) \leq \pi_1 f_1(x)\}} \pi_0 f_0(x)d\mu(x) + \int_{\{\pi_0 f_0(x) > \pi_1 f_1(x)\}} \pi_1 f_1(x)d\mu(x).$$

In our particular case, $BER(g,X,S) = \mathbb{P}(g(X) \neq S)$ when considering $\pi_0 = \pi_1 = \frac{1}{2}$, so we have that

$$g^*(x) = \begin{cases} 1 & if \quad f_0(x) \leq f_1(x) \\ 0 & if \quad f_0(x) > f_1(x) \end{cases}$$

and

$$\min_{g\in\mathcal{G}} BER(g,X,S) = BER(g^*,X,S) = \frac{1}{2}\left[\int_{f_0(x)\leq f_1(x)} f_0(x)d\mu(x) + \int_{f_0(x)>f_1(x)} f_1(x)d\mu(x)\right]$$
$$= \frac{1}{2}\int (f_0\wedge f_1)(x)d\mu(x).$$

This concludes the proof since by definition

$$d_{TV}(\mu_0,\mu_1) = \frac{1}{2}\int |f_0 - f_1|\, d\mu = 1 - \int (f_0\wedge f_1)(x)d\mu(x).$$

□

For the proof of Theorem 4.3.3, we need the following lemma.

Lemma 4.7.1 *Under Assumptions of Theorem 4.3.3, the following bound holds*

$$R(g_B\circ T_S, X) - R(g_B, X, S) \leq 2\mathbb{E}\left[|\eta_S(X) - \eta_S\circ T_S(X)|\right].$$

Proof. We want to be able to control the difference $\inf_{h\in\mathcal{G}} R(h,\tilde{X}) - \inf_{g\in\mathcal{G}} R(g,X,S)$.

To do this, observe that

$$\begin{aligned} R_B(\tilde{X}) - R_B(X,S) &:= \inf_{h\in\mathcal{G}} R(h,\tilde{X}) - \inf_{g\in\mathcal{G}} R(g,X,S) \\ &\leq R(g_B\circ T_S, X) - R(g_B,X,S) = E\left[(2\eta_S(X)-1)(\mathbb{1}_{g_B\circ T_S(X)=0} - \mathbb{1}_{g_B(X,S)=0})\right] \\ &= \mathbb{E}\left[(2\eta_S(X)-1)\mathbb{1}_{g\circ T_S(X)\neq g_B(X,S)}(\mathbb{1}_{g_B\circ T_S(X)\neq 1} - \mathbb{1}_{g_B(X,S)\neq 1})\right], \end{aligned}$$

where the last equality holds because $\left(\mathbb{1}_{g_B\circ T_S(X)\neq 1}\right) - \left(\mathbb{1}_{g_B(X,S)\neq 1}\right) = 0$ if, and only if, both classifiers have the same response $g_B\circ T_S(X) = g_B(X,S)$.

Consider $X = x$ and $S = s$,

- if $g_B(x,s) = 1$, $2\eta_s(x) - 1 \geqslant 0$ and $\mathbb{1}_{g_B(x,s)\neq 1} = 0$. In this situation, we deduce that

$$\mathbb{1}_{g_B\circ T_s(x)\neq g_B(x,s)} = 1 \Leftrightarrow g_B\circ T_s(x) = 0,$$

  and

$$\mathbb{1}_{g_B\circ T_s(x)\neq 1} - \mathbb{1}_{g_B(x,s)\neq 1} = 1.$$

- if $g_B(x,s) = 0$, $2\eta_s(x) - 1 < 0$ and $\mathbb{1}_{g_B(x,s)\neq 1} = 1$. We deduce that

$$\mathbb{1}_{g_B\circ T_s(x)\neq g_B(x,s)} = 1 \Leftrightarrow g_B\circ T_s(x) = 1,$$

  and

$$\mathbb{1}_{g_B\circ T_s(x)\neq 1} - \mathbb{1}_{g_B(x,s)\neq 1} = -1.$$

In any case, the random variable $(2\eta_S(X)-1)\mathbb{1}_{g\circ T_S(X)\neq g_B(X,S)}(\mathbb{1}_{g_B\circ T_S(X)\neq 1} - \mathbb{1}_{g_B(X,S)\neq 1})$ is positive and so it is its expectation

$$R(g_B\circ T_S, X) - R(g_B,X,S) = \mathbb{E}\left[|2\eta_S(X)-1|\,\mathbb{1}_{g\circ T_S(X)\neq g_B(X,S)}\right] \geqslant 0.$$

Moreover, notice that $g_B \circ T_s(x) = \mathbb{1}_{\eta_s \circ T_s(x) > \frac{1}{2}}, forallx, foralls.$ Hence, $g_B \circ T_s(x) \neq g_B(x,s)$ if, and only if, either $\eta_s(x) > \frac{1}{2}$ and $\eta_s \circ T_s(x) < \frac{1}{2}$ or $\eta_s(x) < \frac{1}{2}$ and $\eta_s \circ T_s(x) > \frac{1}{2}$. In both cases,

$$|\eta_s(x) - \eta_s \circ T_s(x)| = \left|\eta_s(x) - \frac{1}{2} + \frac{1}{2} - \eta_s \circ T_s(x)\right| = \left|\eta_s(x) - \frac{1}{2}\right| + \left|\frac{1}{2} - \eta_s \circ T_s(x)\right|,$$

and then it is clear that

$$\left|\eta_s(x) - \frac{1}{2}\right| \leq |\eta_s(x) - \eta_s \circ T_s(x)|, \text{ for all } x, \text{for all } s.$$

In conclusion, the difference between the risk using the Bayes' classifier with the original variable X, S and the modified version $\tilde{X} = T_S(X)$ can be bounded as follows

$$R(g_B \circ T_S, X) - R(g_B, X, S) \leq 2\mathbb{E}\left[|\eta_S(X) - \eta_S \circ T_S(X)|\right].$$

□

Proof of Theorem 4.3.3. First, note that $R(h, \tilde{X}) = R(h, T_S(X)) \leq R(g_B, T_S(X)) = R(g_B \circ T_S, X)$. Thus, it suffices bounding the difference between the minimal risks obtained for the best classifier with input data (X, S), called g_B, and the risk obtained with this classification rule using the input data $\tilde{X}$

$$\begin{aligned} &R(g_B \circ T_S, X) - R(g_B, X, S) \leq 2\mathbb{E}_{(X,S)}\left[|\eta_S(X) - \eta_S \circ T_S(X)|\right] \\ &= 2\left[\mathbb{P}(S=0)\mathbb{E}_X\left[|\eta_0(X) - \eta_0 \circ T_0(X)| \mid S=0\right] + \mathbb{P}(S=1)\mathbb{E}_X\left[|\eta_1(X) - \eta_1 \circ T_1(X)| \mid S=1\right]\right] \\ &= 2\sum_{s=0,1} \pi_s \mathbb{E}_X\left[|\eta_s(X) - \eta_s \circ T_s(X)| \mid S=s\right]. \end{aligned}$$

Moreover, by the Lipschitz condition and noting that $a + b \leq 2^{\frac{1}{2}}(a^2+b^2)^{\frac{1}{2}}, \ forall\ a, b \in \mathbb{R}$, we can write

$$\begin{aligned} R(g_B \circ T_S, X) - R(g_B, X, S) &\leq 2\sum_{s=0,1} \pi_s K_s \mathbb{E}_X\left[\|X - T_s(X)\| \mid S=s\right] \\ &\leq 2\sqrt{2}K\left(\sum_{s=0,1} \pi_s^2 \left(\mathbb{E}_X\left[\|X - T_s(X)\| \mid S=s\right]\right)^2\right)^{\frac{1}{2}}, \end{aligned}$$

where $K = \max\{K_0, K_1\}$. Finally, the Cauchy-Schwarz inequality gives

$$\begin{aligned} R(g_B \circ T_S, X) - R(g_B, X, S) &\leq 2\sqrt{2}K\left(\sum_{s=0,1} \pi_s^2 \mathbb{E}_X\left[\|X - T_s(X)\|^2 \mid S=s\right]\right)^{\frac{1}{2}} \\ &= 2\sqrt{2}K\left(\sum_{s=0,1} \pi_s^2 \mathcal{W}_2^2(\mu_s, \mu_{s\sharp}T_s)\right)^{\frac{1}{2}} \leq 2\sqrt{2}K\left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_{s\sharp}T_s)\right)^{\frac{1}{2}}. \end{aligned}$$

□

4.7.2 Application on a real dataset

To ilustrate the performance of the repairing procedures in Section 3, we consider the *Adult Income* data set (available at `https://archive.ics.uci.edu/ml/datasets/adult`). It contains 29.825 instances consisting in the values of 14 attributes, 6 numeric and 8 categorical, and a categorization of each person as having an income of more or less than $50,000$\$ per year. This attribute will be the target variable in the study. In the following, we estimate the Disparate Impact using its empirical counterpart and provide a confidence interval which was established in Besse et al. [2018b]. Among the rest of the categorical attributes, we focus on the sensitive attribute *Gender* ("male"or "female") to be the potentially protected. As the repairing procedures work only with the numerical attributes, to check their effectiveness we will follow the next steps:

1. Split the data set into the test and the learning sample using the ratio 2.500 / 27.325.
2. Train the classifiers based on logistic regression and random forests using the five numerical variables: *Age, Education Level, Capital Gain, Capital Loss* and *Worked hours per week*.
3. Predict the target for the test sample with the built model and compute the misclassification error of each rule.
4. Apply the repair procedure to the test sample described by the numerical variables.
5. Predict the target for the repaired data set with the built model and compute the misclassification error again.

In Table 4.2 a summary of the performance of the two classification rules considered is presented. With a confidence of 95%, we can say that the logit classifier has Disparate Impact at level 0.555 and the Random Forests at 0.54, with respect to Gender. Hence, both rules are committing discrimination with respect to this sensitive variable. Now we will see how the repairing procedures studied in section 4.3 help in blurring the protected variable.

In Table 4.3 we can see that in the experiments with procedure **(A)** the estimated value for DI is not exactly 1, as we have already anticipated. On the other hand, procedure **(B)** manages to change the data in such a way that both classification rules attain Statistical Parity. Moreover, the error in the classification done with the repaired data sets is smaller when using procedure **(B)** in the two cases. In Feldman et al. [2015], they propose a generalization to higher dimension by computing the repairing procedure for each attribute. This procedure is denoted in the table with the letter **(C)**. We see that the error is smaller than with **(A)** but still much bigger than with **(B)**. Moreover, the estimated level of Disparate Impact is not 1 but it is closer to the Statistical Parity than with procedure **(A)**.

Finally, we present some results of the performance of the Geometric and Random Repairs. Left part of Figures 4.9 and 4.10 represent the evolution of the estimated Disparate Impact with the amount of repair $0 \leq \lambda \leq 1$, while the right part show the evolution with λ of the error in the classification done from the modified data set. For the experiments concerning the Random Repair procedure (denoted RR in the figures) we have repeated it 100 times, and then we have computed the mean of the simulations. Clearly, the level of DI reached is higher with the Random Repair for the logit rule. For the random forest procedure since the rule is not linear, the difference is not as high and Disparate Impacts have similar behaviors. Yet for larger amount of repair the gap between the two different kinds of repair increases at the advantage of the Geometric Repair.

Moreover, the error in the prediction from the new data modified with this procedure is smaller than with the Geometric Repair. We note that the amount of repair necessary to

achieve a confidence interval for DI at level 0.8 for the logit rule is 0.3 with the Random Repair, which entails an error of 0.2068; and 0.55 with the Geometric Repair, which entails an error of 0.2136. In the case of the random forests rule, this value is 0.5 for both but the error is 0.1927 with the Random Repair; and 0.2076 with the Geometric Repair.

Table 4.2 – Performance and Disparate Impact with respect to the protected variable Gender.

Statistical Model	Error	$\hat{DI}$	CI 95%
Logit	0.2064	0.496	$(0.437, 0.555)$
Random Forests	0.168	0.484	$(0.429, 0.54)$

Table 4.3 – Repairing procedures and Disparate impact of the rules with the modified dataset

Statistical Model	Repair	Error	Difference	$\hat{DI}$	CI 95%
Logit	**(A)**	0.218	0.0116	0.937	$(0.841, 1.033)$
Logit	**(B)**	0.2077	0.00128	1	$(0.905, 1.095)$
Logit	**(C)**	0.2132	0.0068	0.94	$(0.842, 1.038)$
Random Forests	**(A)**	0.2272	0.0592	1.1	$(0.976, 1.223)$
Random Forests	**(B)**	0.2045	0.0365	1	$(0.886, 1.114)$
Random Forests	**(C)**	0.2152	0.0472	1.091	$(0.978, 1.203)$

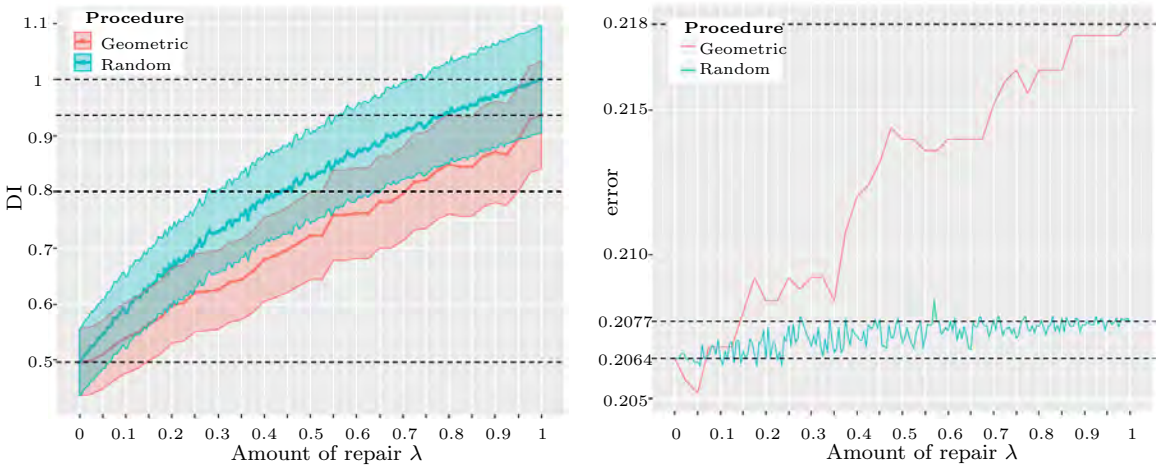


Figure 4.9 – CI at level 95% for DI (left) and error (right) of the classifier logit with respect to Gender and the data repaired by the Geometric and Random Repair

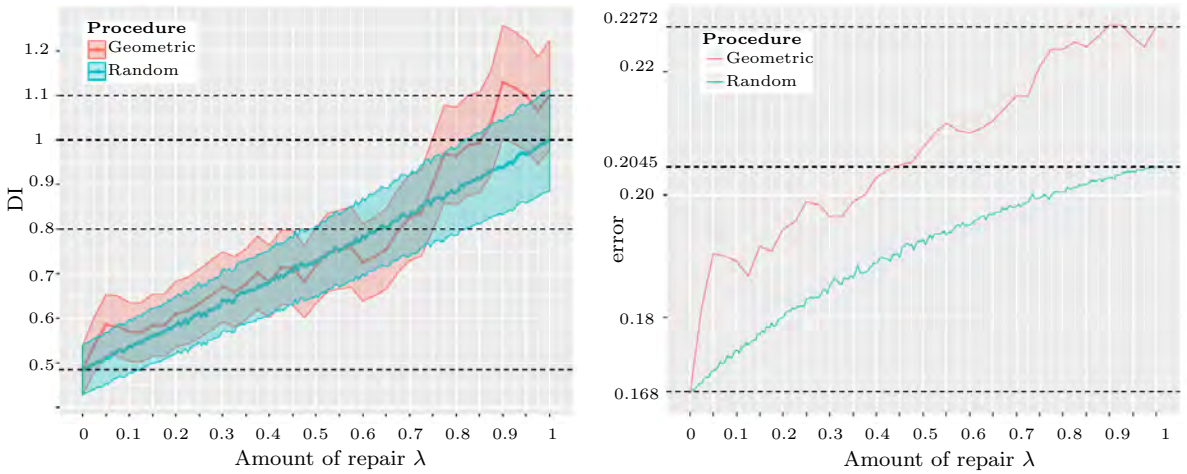


Figure 4.10 – CI at level 95% for DI (left) and error (right) of the classifier random forests with respect to Gender and the data repaired by the Geometric and Random Repair

4.8 Appendix B to Chapter 4

4.8.1 Quantifying the loss when predicting with LASSO from the repaired data through scale-location models

In this appendix we describe how imposing fairness constraints affects the quality of the prediction when considering LASSO estimation.

We observe $(X_1, S_1, Y_1), \ldots, (X_n, S_n, Y_n)$ i.i.d. from the random vector (X, S, Y), where $Y \in \mathbb{R}$, $X \in \mathbb{R}^p$ and $S \in \{0, 1\}$ correspond respectively to response, usable and protected attribute, as usual. Consider the following model

$$Y = (I_n - S)\mathbb{X}\alpha + S\mathbb{X}\beta + \varepsilon, \tag{4.8.1}$$

where $Y = [Y_1, \ldots, Y_n]^T$, $\mathbb{X} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{bmatrix}$ is the design matrix, $\alpha = [\alpha_0, \alpha_1, \ldots, \alpha_p]^T$, $\beta = [\beta_0, \beta_1, \ldots, \beta_p]^T$, $S = diag(S_1, \ldots, S_n)$ and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$. We will assume that model (4.8.1) holds exactly, with some true parameter values α^* and β^* and we will denote $f^* = (I_n - S)\mathbb{X}\alpha^* + S\mathbb{X}\beta^*$. Thus, the true model can be written $Y = f^* + \varepsilon$.

Now we want to predict Y using a transformation of X that depends on the protected group S, say $\tilde{X} = T_S(X)$, such that the impredictability of the protected attribute from the modified data is ensured, which means that

$$\mathcal{L}(\tilde{X} \mid S = 0) = \mathcal{L}(\tilde{X} \mid S = 1). \tag{4.8.2}$$

In particular, we will consider location-scale transformations. Thus, we will write for the transformed design matrix

$$\tilde{\mathbb{X}} = (I_n - S)\mathbb{X}M_0 + S\mathbb{X}M_1, \tag{4.8.3}$$

where

$$M_s = \left[\begin{array}{c|c} 1 & b_s^T \\ 1\times 1 & 1\times p \\ \hline 0 & A_s \\ p\times 1 & p\times p \end{array}\right],\ A_s = diag(a_{s1},\ldots,a_{sp}),\ b_s = [b_{s1},\ldots,b_{sp}]^T \tag{4.8.4}$$

and b_s, A_s are the location and scale parameters, respectively, of the transformation $T_s,\ s = 0, 1$. More precisely, we have that $\tilde{\mathbb{X}} = \left(\tilde{\mathbb{X}}_{ij}\right)_{1\le i\le n, 1\le j\le p+1}$ with

$$\begin{cases} \tilde{\mathbb{X}}_{ij} = b_{0j} + a_{0j}X_{ij},\ 2\le j\le p+1,\ \text{if } S_i = 0 \\ \tilde{\mathbb{X}}_{ij} = b_{1j} + a_{1j}X_{ij},\ 2\le j\le p+1,\ \text{if } S_i = 1 \\ \tilde{\mathbb{X}}_{i1} = 1,\ i = 1,\ldots,n \end{cases} \tag{4.8.5}$$

In order to quantify the loss in the prediction of Y from the modified data, we will consider the Lasso

$$\hat{\tilde{\beta}} := argmin_{\tilde{\beta}\in\mathbb{R}^{p+1}}\left\{\frac{\|Y-\tilde{\mathbb{X}}\tilde{\beta}\|_2^2}{n} + \lambda\|\tilde{\beta}\|_1\right\},$$

where the coefficient $\tilde{\beta}$ does not depend on the value of $S = s$. By definition of the Lasso, for every $\beta\in\mathbb{R}^{p+1}$ it holds that

$$\frac{\|Y-\tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 \le \frac{\|Y-\tilde{\mathbb{X}}\beta\|_2^2}{n} + \lambda\|\beta\|_1. \tag{4.8.6}$$

If we write

$$\|Y-\tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2 = \|f^*+\varepsilon-\tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2 = \|f^*-\tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2 + \|\varepsilon\|_2^2 + 2\varepsilon^T(f^*-\tilde{\mathbb{X}}\hat{\tilde{\beta}}),$$

and the same for β in the right-hand side of (4.8.6), then we have the following basic inequality

$$\begin{aligned}\frac{\|f^*-\tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 &\le \frac{\|f^*-\tilde{\mathbb{X}}\beta\|_2^2}{n} + \frac{2\varepsilon^T(f^*-\tilde{\mathbb{X}}\beta)}{n} - \frac{2\varepsilon^T(f^*-\tilde{\mathbb{X}}\hat{\tilde{\beta}})}{n} + \lambda\|\beta\|_1 \\ &= \frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}}-\beta)}{n} + \lambda\|\beta\|_1 + \frac{\|f^*-\tilde{\mathbb{X}}\beta\|_2^2}{n}.\end{aligned}$$

The first term represents the random part where the measurement error plays a role. This part can be easily bounded in terms of the l_1-norm of the parameters involved as follows

$$\begin{aligned}\left|\frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}}-\beta)}{n}\right| &= \left|\frac{2\varepsilon^T((I_n-S)\mathbb{X}M_0 + S\mathbb{X}M_1)(\hat{\tilde{\beta}}-\beta)}{n}\right| \\ &\le \left|\frac{2\varepsilon^T(I_n-S)\mathbb{X}M_0(\hat{\tilde{\beta}}-\beta)}{n}\right| + \left|\frac{2\varepsilon^TS\mathbb{X}M_1(\hat{\tilde{\beta}}-\beta)}{n}\right| \\ &\le \max_{1\le j\le p+1}\frac{2|(\varepsilon^T(I_n-S)\mathbb{X})_{.j}|}{n}\|M_0(\hat{\tilde{\beta}}-\beta)\|_1 \\ &\quad + \max_{1\le j\le p+1}\frac{2|(\varepsilon^TS\mathbb{X})_{.j}|}{n}\|M_1(\hat{\tilde{\beta}}-\beta)\|_1.\end{aligned}$$

The idea of the penalty in the Lasso is that it should be chosen to control this random part. Let us therefore introduce the sets

$$\mathcal{J}_0 := \left\{ \max_{1\le j\le p+1} 2|(\varepsilon^T(I_n - S)\mathbb{X})_{.j}| \le \lambda_0 \right\}$$
$$\mathcal{J}_1 := \left\{ \max_{1\le j\le p+1} 2|(\varepsilon^T S\mathbb{X})_{.j}| \le \lambda_0 \right\},$$

where we assume $\lambda \ge k\lambda_0$, for some constant $k > 0$ that will be fixed later, to make sure that on $\mathcal{J} := \mathcal{J}_0 \cap \mathcal{J}_1$ we can get rid of the random part of the problem.

We will show next that for a suitable value of λ_0, the set $\mathcal{J}$ has large probability. Let us denote $\hat{\Sigma} = \frac{\mathbb{X}^T\mathbb{X}}{n}$ and $\hat{\sigma}_j^2 := \hat{\Sigma}_{jj},\ j = 1, \dots p+1$, its diagonal elements

$$\hat{\Sigma} = \left[\begin{array}{c|c} 1 & \bar{X}_n^T \\ {\scriptstyle 1\times 1} & {\scriptstyle 1\times p} \\ \hline \bar{X}_n & X^T X \\ {\scriptstyle p\times 1} & {\scriptstyle p\times p} \end{array}\right] \tag{4.8.7}$$

Lemma 4.8.1 *Suppose that $\hat{\sigma}_j^2 = 1$ for all j. Then we have for all $t > 0$, and for*

$$\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log(p+1)}{\min\{n_0, n_1\}}},$$
$$P(\mathcal{J}) \ge 1 - 4\exp(-t^2/2).$$

Proof of Lemma 4.8.1. As $\hat{\sigma}_j^2 = 1$, the random variables

$$V_j^0 := \frac{(\varepsilon^T(I_n - S)\mathbb{X})_{.j}}{n_0\sigma^2} \sim N(0, \sigma_0^2)$$
$$V_j^1 := \frac{(\varepsilon^T S\mathbb{X})_{.j}}{n_1\sigma^2} \sim N(0, \sigma_1^2),$$

with variances $\sigma_0^2, \sigma_1^2 \le 1$. Hence, for $s = 0, 1$,

$$P\left(\max_{1\le j\le p+1} |V_j^s| > \sqrt{t^2 + 2\log(p+1)}\right) \le 2(p+1)\exp(\frac{-t^2 + 2\log(p+1)}{2}) = 2\exp(-t^2/2), \tag{4.8.8}$$

and consequently,

$$P\left(\max_{1\le j\le p+1} \frac{2|(\varepsilon^T(I_n - S)\mathbb{X})_{.j}|}{n_0} > \frac{2\sigma}{\sqrt{n_0}}\sqrt{t^2 + 2\log(p+1)}\right) \le 2\exp(-t^2/2) \tag{4.8.9}$$
$$P\left(\max_{1\le j\le p+1} \frac{2|(\varepsilon^T S\mathbb{X})_{.j}|}{n_1} > \frac{2\sigma}{\sqrt{n_1}}\sqrt{t^2 + 2\log(p+1)}\right) \le 2\exp(-t^2/2). \tag{4.8.10}$$

We deduce that if we take $\lambda_0 = 2\sigma\sqrt{\frac{t^2+2\log(p+1)}{\min\{n_0,n_1\}}}$ we have that $P(\mathcal{J}) \ge 1 - 4\exp(-t^2/2)$.

□

If we denote for $s = 0, 1$, $\hat{\pi}_s = \frac{n_s}{n}$, then on the set $\mathcal{J}$ we have for every $\beta \in \mathbb{R}^{p+1}$

$$\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 \le \lambda\left(\frac{\hat{\pi}_0}{k}\|M_0(\hat{\tilde{\beta}} - \beta)\|_1 + \frac{\hat{\pi}_1}{k}\|M_1(\hat{\tilde{\beta}} - \beta)\|_1 + \|\beta\|_1\right) + \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n}. \tag{4.8.11}$$

Let us now consider $\beta^0 := argmin_{\tilde{\beta}\in\mathbb{R}^{p+1}}\left\{\frac{\|Y-\tilde{\mathbb{X}}\tilde{\beta}\|_2^2}{n}\right\}$ and assume that $f^* = \mathbb{X}\beta^0$, which means that the true response is linear. Then, the basic inequality becomes

$$\frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 \leq \frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - \beta)}{n} + \lambda\|\beta\|_1 + \frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\beta\|_2^2}{n},$$

for all $\beta \in \mathbb{R}^{p+1}$.

For the proof of Corollary 4.8.3, the following observation is needed:

Remark 4.8.2 (Bound for the norm of the transformed parameters)

$$M_s\hat{\tilde{\beta}} = \begin{bmatrix} 1 & b_{s1} & b_{s2} & \dots & b_{sp} \\ 0 & a_{s1} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & a_{sp} \end{bmatrix} \begin{bmatrix} \hat{\tilde{\beta}}_0 \\ \hat{\tilde{\beta}}_1 \\ \vdots \\ \hat{\tilde{\beta}}_p \end{bmatrix} = \begin{bmatrix} \hat{\tilde{\beta}}_0 + \sum_{j=1}^p b_{s,j}\hat{\tilde{\beta}}_j \\ a_{s,1}\hat{\tilde{\beta}}_1 \\ \vdots \\ a_{s,p}\hat{\tilde{\beta}}_p \end{bmatrix}$$

$$\begin{aligned} \Rightarrow \|M_s\hat{\tilde{\beta}}\|_1 &= |\hat{\tilde{\beta}}_0 + \sum_{j=1}^p b_{s,j}\hat{\tilde{\beta}}_j| + |\sum_{j=1}^p a_{s,j}\hat{\tilde{\beta}}_j| \leq |\hat{\tilde{\beta}}_0| + \sum_{j=1}^p |b_{s,j}\hat{\tilde{\beta}}_j| + \sum_{j=1}^p |a_{s,j}\hat{\tilde{\beta}}_j| \\ &\leq |\hat{\tilde{\beta}}_0| + \|b_s\|_\infty(\|\hat{\tilde{\beta}}\|_1 - |\hat{\tilde{\beta}}_0|) + \|a_s\|_\infty(\|\hat{\tilde{\beta}}\|_1 - |\hat{\tilde{\beta}}_0|) \\ &= (\|b_s\|_\infty + \|a_s\|_\infty)\|\hat{\tilde{\beta}}\|_1 + (1 - \|b_s\|_\infty - \|a_s\|_\infty)|\hat{\tilde{\beta}}_0| = c_s\|\hat{\tilde{\beta}}\|_1 + (1 - c_s)|\hat{\tilde{\beta}}_0|, \end{aligned}$$

where $c_s = \|b_s\|_\infty + \|a_s\|_\infty$, $\|b_s\|_\infty = \max_{j=1,\dots,p}|b_{s,j}|$, *and* $\|a_s\|_\infty = \max_{j=1,\dots,p}|a_{s,j}|$

Corollary 4.8.3 *Assume than* $\hat{\sigma}_j^2 = 1$ *for all* j. *For some* $t > 0$, *let the regularization parameter be*

$$\lambda = 2k\hat{\sigma}\sqrt{\frac{t^2 + 2\log(p+1)}{\min\{n_0, n_1\}}},$$

where $\hat{\sigma}$ *is an estimator of* σ *Then with probability at least* $1 - \alpha$, *where*

$$\alpha := 4\exp(-t^2/2) + P(\hat{\sigma} \leq \sigma),$$

we have

$$\begin{aligned} \frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} \leq \lambda \Big[&(1 + \frac{\nu}{k})\left(\hat{\pi}_0\|M_0^{-1}\beta^0\|_1 + \hat{\pi}_1\|M_1^{-1}\beta^0\|_1\right) \\ &+(1 - \frac{\nu}{k})\left(|\hat{\tilde{\beta}}_0| + \hat{\pi}_0|(M_0^{-1}\beta^0)_0| + \hat{\pi}_1|(M_1^{-1}\beta^0)_0|\right)\Big] \\ &+ \hat{\pi}_0\frac{\|S\mathbb{X}(M_0^{-1}\beta^0 - M_1^{-1}\beta^0)\|_2^2}{n} + \hat{\pi}_1\frac{\|(I_n - S)\mathbb{X}(M_0^{-1}\beta^0 - M_1^{-1}\beta^0)\|_2^2}{n}. \end{aligned}$$

Proof of Corollary 4.8.3. In particular, for β^0 and on the set $\mathcal{J}$ we have

$$\begin{aligned} \frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 &\leq \frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - \beta^0)}{n} + \lambda\|\beta^0\|_1 + \frac{\|(\mathbb{X} - \tilde{\mathbb{X}})\beta^0\|_2^2}{n} \\ &\leq \lambda\left(\frac{\hat{\pi}_0}{k}\|M_0(\hat{\tilde{\beta}} - \beta^0)\|_1 + \frac{\hat{\pi}_1}{k}\|M_1(\hat{\tilde{\beta}} - \beta^0)\|_1\right) + \lambda\|\beta^0\|_1 + \frac{\|(\mathbb{X} - \tilde{\mathbb{X}})\beta^0\|_2^2}{n}, \end{aligned}$$

In particular, for $M_0^{-1}\beta^0$ and $M_1^{-1}\beta^0$ it holds that

$$\frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 \leq \frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - M_0^{-1}\beta^0)}{n} + \lambda\|M_0^{-1}\beta^0\|_1 + \frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}M_0^{-1}\beta^0\|_2^2}{n} \quad (4.8.12)$$

$$\frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 \leq \frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - M_1^{-1}\beta^0)}{n} + \lambda\|M_1^{-1}\beta^0\|_1 + \frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}M_1^{-1}\beta^0\|_2^2}{n}. \quad (4.8.13)$$

Then, adding up expressions in both inequalities with weights $\hat{\pi_0}$ and $\hat{\pi_1}$, respectively, we have that:

$$\begin{aligned}\frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 \leq \hat{\pi}_0 &\left(\frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - M_0^{-1}\beta^0)}{n} + \lambda\|M_0^{-1}\beta^0\|_1 + \frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}M_0^{-1}\beta^0\|_2^2}{n}\right) \\ &+ \hat{\pi}_1\left(\frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - M_1^{-1}\beta^0)}{n} + \lambda\|M_1^{-1}\beta^0\|_1 + \frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}M_1^{-1}\beta^0\|_2^2}{n}\right).\end{aligned}$$

From the definition of the transformation $\tilde{\mathbb{X}}$, we can simplify the expression on the right-hand side of the inequality above by observing that:

1.
$$\begin{aligned}&\hat{\pi}_0\frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}M_0^{-1}\beta^0\|_2^2}{n} + \hat{\pi}_1\frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}M_1^{-1}\beta^0\|_2^2}{n} \\ &= \hat{\pi}_0\frac{\|\mathbb{X}\beta^0 - (I_n - S)\mathbb{X}\beta^0 - S\mathbb{X}M_1M_0^{-1}\beta^0\|_2^2}{n} + \hat{\pi}_1\frac{\|\mathbb{X}\beta^0 - (I_n - S)\mathbb{X}M_0M_1^{-1}\beta^0 - S\mathbb{X}\beta^0\|_2^2}{n} \\ &= \hat{\pi}_0\frac{\|S\mathbb{X}(\beta^0 - M_1M_0^{-1}\beta^0)\|_2^2}{n} + \hat{\pi}_1\frac{\|(I_n - S)\mathbb{X}(\beta^0 - M_0M_1^{-1}\beta^0)\|_2^2}{n} \\ &= \hat{\pi}_0\frac{\|S\mathbb{X}(M_0^{-1}\beta^0 - M_1^{-1}\beta^0)\|_2^2}{n} + \hat{\pi}_1\frac{\|(I_n - S)\mathbb{X}(M_0^{-1}\beta^0 - M_1^{-1}\beta^0)\|_2^2}{n}\end{aligned}$$

2. On the set $\mathcal{J}$, if $\lambda \geq k\lambda_0$, reasoning as before

$$\begin{aligned}\left|\frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - M_0^{-1}\beta^0)}{n}\right| &\leq \left|\frac{2\varepsilon^T(I_s - S)\mathbb{X}(M_0\hat{\tilde{\beta}} - \beta^0)}{n}\right| + \left|\frac{2\varepsilon^TS\mathbb{X}(M_1\hat{\tilde{\beta}} - M_1M_0^{-1}\beta^0)}{n}\right| \\ &\leq \lambda\left(\frac{\hat{\pi}_0}{k}\|M_0\hat{\tilde{\beta}} - \beta^0\|_1 + \frac{\hat{\pi}_1}{k}\|M_1\hat{\tilde{\beta}} - M_1M_0^{-1}\beta^0\|_1\right) \\ &\leq \lambda\left(\frac{\hat{\pi}_0}{k}\|M_0(\hat{\tilde{\beta}} - M_0^{-1}\beta^0)\|_1 + \frac{\hat{\pi}_1}{k}\|M_1(\hat{\tilde{\beta}} - M_0^{-1}\beta^0)\|_1\right) \\ &\leq \lambda\left(\frac{\nu}{k}\|\hat{\tilde{\beta}} - M_0^{-1}\beta^0\|_1 + (1 - \frac{\nu}{k})|(\hat{\tilde{\beta}} - M_0^{-1}\beta^0)_0|\right)\end{aligned}$$

and similarly

$$\begin{aligned}\left|\frac{2\varepsilon^T\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - M_1^{-1}\beta^0)}{n}\right| &\leq \left|\frac{2\varepsilon^T(I_s - S)\mathbb{X}(M_0\hat{\tilde{\beta}} - M_0M_1^{-1}\beta^0)}{n}\right| + \left|\frac{2\varepsilon^TS\mathbb{X}(M_1\hat{\tilde{\beta}} - \beta^0)}{n}\right| \\ &\leq \lambda\left(\frac{\hat{\pi}_0}{k}\|M_0\hat{\tilde{\beta}} - M_0M_1^{-1}\beta^0\|_1 + \frac{\hat{\pi}_1}{k}\|M_1\hat{\tilde{\beta}} - \beta^0\|_1\right) \\ &\leq \lambda\left(\frac{\hat{\pi}_0}{k}\|M_0(\hat{\tilde{\beta}} - M_1^{-1}\beta^0)\|_1 + \frac{\hat{\pi}_1}{k}\|M_1(\hat{\tilde{\beta}} - M_1^{-1}\beta^0)\|_1\right) \\ &\leq \lambda\left(\frac{\nu}{k}\|\hat{\tilde{\beta}} - M_1^{-1}\beta^0\|_1 + (1 - \frac{\nu}{k})|(\hat{\tilde{\beta}} - M_1^{-1}\beta^0)_0|\right),\end{aligned}$$

where in the last inequality $\nu = \hat{\pi}_0c_0 + \hat{\pi}_1c_1$, with values c_s defined in Remark 4.8.2.

So, gathering observations 1 and 2, we deduce that

$$\begin{aligned}\frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 \leq \lambda \Big[&\frac{\nu}{k}\left(\hat{\pi}_0\|\hat{\tilde{\beta}} - M_0^{-1}\beta^0\|_1 + \hat{\pi}_1\|\hat{\tilde{\beta}} - M_1^{-1}\beta^0\|_1\right)\\ &+(1-\frac{\nu}{k})\left(\hat{\pi}_0|(\hat{\tilde{\beta}} - M_0^{-1}\beta^0)_0| + \hat{\pi}_1|(\hat{\tilde{\beta}} - M_1^{-1}\beta^0)_0|\right)\\ &+\hat{\pi}_0\,\|M_0^{-1}\beta^0\|_1 + \hat{\pi}_1\|M_1^{-1}\beta^0\|_1\Big]\\ &+\hat{\pi}_0\frac{\|S\mathbb{X}(M_0^{-1}\beta^0 - M_1^{-1}\beta^0)\|_2^2}{n} + \hat{\pi}_1\frac{\|(I_n - S)\mathbb{X}(M_0^{-1}\beta^0 - M_1^{-1}\beta^0)\|_2^2}{n}.\end{aligned}$$

Applying the triangle inequality we finally obtain

$$\begin{aligned}\frac{\|\mathbb{X}\beta^0 - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}\|_1 \leq \lambda \Big[&\frac{\nu}{k}\|\hat{\tilde{\beta}}\|_1 + (1+\frac{\nu}{k})\left(\hat{\pi}_0\|M_0^{-1}\beta^0\|_1 + \hat{\pi}_1\|M_1^{-1}\beta^0\|_1\right)\\ &+(1-\frac{\nu}{k})\left(|\hat{\tilde{\beta}}_0| + \hat{\pi}_0|(M_0^{-1}\beta^0)_0| + \hat{\pi}_1|(M_1^{-1}\beta^0)_0|\right)\Big]\\ &+\hat{\pi}_0\frac{\|S\mathbb{X}(M_0^{-1}\beta^0 - M_1^{-1}\beta^0)\|_2^2}{n} + \hat{\pi}_1\frac{\|(I_n - S)\mathbb{X}(M_0^{-1}\beta^0 - M_1^{-1}\beta^0)\|_2^2}{n}.\end{aligned}$$

□

4.8.1.1 Assuming sparsity in the parameters

Let us consider now a fixed $\beta \in \mathbb{R}^{p+1}$ and suppose that only a few of its components, say $0 < s \leq p+1$, are non-zero. For an index set $Z \subset \{0, \ldots, p\}$ we will write

$$\begin{aligned}\beta_{j,Z} &:= \beta_j I(j \in Z)\\ \beta_{j,Z^c} &:= \beta_j I(j \notin Z).\end{aligned}$$

Thus $\beta_Z := [\beta_{0,Z}, \ldots, \beta_{p,Z}]$ has nule entries at least for the indexes outside the set Z. Similarly β_{Z^c} has at least $(p+1) - s$ zeroes in the positions in Z. Clearly, $\beta = \beta_Z + \beta_{Z^c}$. Moreover, we will write $Z(\beta) := \{j \in \{0, \ldots, p\}/\beta_j \neq 0\}$ the set of indexes that correspond to non-zero entries in β, so $|Z(\beta)| = s$.

Lemma 4.8.4 *For every $\beta \in \mathbb{R}^{p+1}$ we have on $\mathcal{J}$, with $\lambda \geq k\lambda_0$,*

$$\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1 \leq \lambda\left(1 + \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 + \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n}.$$

Proof of Lemma 4.8.4. By the second tringle inequality, it is clear that

$$\begin{aligned}\|\hat{\tilde{\beta}}\|_1 &= \|\hat{\tilde{\beta}}_{Z(\beta)}\|_1 + \|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1 = \|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)} + \beta_{Z(\beta)}\|_1 + \|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1\\ &\geq \|\beta_{Z(\beta)}\|_1 - \|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 + \|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1.\end{aligned}$$

Thus, taking this into the left-hand side of (4.8.11) and noting that by Remark

$$\|M_s(\hat{\tilde{\beta}} - \beta)\|_1 \leq c_s\|\hat{\tilde{\beta}} - \beta\|_1 + (1 - c_s)|(\hat{\tilde{\beta}} - \beta)_0|$$

we obtain

$$\begin{aligned}\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1 &\leq \lambda\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 - \lambda\|\beta_{Z(\beta)}\|_1 \\ &+ \lambda\left(\frac{\hat{\pi}_0}{k}\|M_0(\hat{\tilde{\beta}} - \beta)\|_1 + \frac{\hat{\pi}_1}{k}\|M_1(\hat{\tilde{\beta}} - \beta)\|_1 + \|\beta_{Z(\beta)}\|_1\right) + \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n} \\ &\leq \lambda\left(1 + \frac{c_0\hat{\pi}_0 + c_1\hat{\pi}_1}{k}\right)\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 + \lambda\left(\frac{c_0\hat{\pi}_0 + c_1\hat{\pi}_1}{k}\right)\|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1 \\ &- \lambda\frac{c_0\hat{\pi}_0 + c_1\hat{\pi}_1}{k}|(\hat{\tilde{\beta}} - \beta)_0| + \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n}\end{aligned}$$

which, if we denote $\nu = c_0\hat{\pi}_0 + c_1\hat{\pi}_1 > 0$, finally gives us

$$\begin{aligned}\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1 &\leq \lambda\left(1 + \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 \\ &- \lambda\frac{\nu}{k}|(\hat{\tilde{\beta}} - \beta)_0| + \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n} \\ &\leq \lambda\left(1 + \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 + \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n}.\end{aligned}$$

□

Remark 4.8.5 (Compatibility conditions) *We say that the compatibility condition is met for the set Z if for some constant $\phi(Z) > 0$ and for all β satisfying $\|\beta_{Z^c}\|_1 \leq \frac{\left(2+\frac{\nu}{k}\right)}{\left(1-\frac{\nu}{k}\right)}\|\beta_Z\|_1$, it holds that*

$$\|\beta_Z\|_1^2 \leq \frac{|Z|}{\phi(Z)^2}\frac{\|\tilde{\mathbb{X}}\beta\|_2^2}{n}.$$

Theorem 4.8.6 *Assume than $\hat{\sigma}_j^2 = 1$ for all j. For some $t > 0$, let the regularization parameter be*

$$\lambda = 2k\hat{\sigma}\sqrt{\frac{t^2 + 2\log(p+1)}{\min\{n_0, n_1\}}},$$

where $\hat{\sigma}$ is an estimator of σ. If for a given $\beta \in \mathbb{R}^{p+1}$ the set $Z(\beta)$ satisfies the compatibility condition, then with probability at least $1 - \alpha$, where

$$\alpha := 4\exp(-t^2/2) + P(\hat{\sigma} \leq \sigma),$$

we have

$$\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}} - \beta\|_1 \leq \frac{3s}{2\phi^2(\beta)} + \lambda^2\left(\frac{9s}{\phi^2(\beta)} + 12(1 - \frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}\right).$$

Proof of Theorem 4.8.6. We notice that on the right-hand side, we could have two different cases depending on which of the two expressions related to β is the larger one:

i) $\lambda\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 \geq \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n}$

ii) $\lambda\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 \leq \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n}$

In the first case,

$$\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1 \leq \lambda\left(2 + \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1,$$

which in particular implies that

$$\|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1 \leq \frac{\left(2 + \frac{\nu}{k}\right)}{\left(1 - \frac{\nu}{k}\right)}\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1. \tag{4.8.14}$$

$$\begin{aligned}
&\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}} - \beta\|_1 \\
=&\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\left(\|\hat{\tilde{\beta}}_{Z^c(\beta)}\|_1 + \|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1\right) \\
\leq&\lambda\left(1 + \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 + \frac{\|f^* - \tilde{\mathbb{X}}\beta\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 \\
\leq&3\lambda\|\hat{\tilde{\beta}}_{Z(\beta)} - \beta_{Z(\beta)}\|_1 \leq \frac{3\lambda\sqrt{s}}{\phi(\beta)}\frac{\|\tilde{\mathbb{X}}(\hat{\tilde{\beta}} - \beta)\|_2}{\sqrt{n}},
\end{aligned}$$

where the last inequality follows from applying the compatibility condition to the set $Z(\beta)$, since $\hat{\tilde{\beta}}_{Z^c(\beta)} - \beta_{Z^c(\beta)} = \hat{\tilde{\beta}}_{Z^c(\beta)}$ satisfies (4.8.14). Now, the triangle inequality implies

$$\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}} - \beta\|_1 \leq \frac{3\lambda\sqrt{s}}{\phi(\beta)}\frac{\|\tilde{\mathbb{X}}\hat{\tilde{\beta}} - f^*\|_2}{\sqrt{n}} + \frac{3\lambda\sqrt{s}}{\phi(\beta)}\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2}{\sqrt{n}}.$$

Using inequalities $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$ and $\left(\frac{a}{\sqrt{12(1-\frac{\nu}{k})}}\right)\left(\sqrt{12(1-\frac{\nu}{k})}b\right) \leq \frac{a^2}{24(1-\frac{\nu}{k})} + 6(1-\frac{\nu}{k})b^2$ for the first and second, respectively, terms in the right-hand side,

$$\begin{aligned}
&\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}} - \beta\|_1 \\
&\leq \frac{9\lambda^2 s}{2\phi^2(\beta)} + \frac{\|\tilde{\mathbb{X}}\hat{\tilde{\beta}} - f^*\|_2^2}{2n} + \frac{9s}{24(1-\frac{\nu}{k})\phi^2(\beta)} + 6\lambda^2(1-\frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n},
\end{aligned}$$

and finally

$$\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + 2\lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}} - \beta\|_1 \leq \frac{9\lambda^2 s}{\phi^2(\beta)} + \frac{9s}{12(1-\frac{\nu}{k})\phi^2(\beta)} + 12\lambda^2(1-\frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}.$$

If we choose $k > 0$ such that $2\left(1 - \frac{\nu}{k}\right) \geq 1$, that is $\Leftrightarrow k \geq 2\nu$, then from Lemma 4.8.1 we know that for $t > 0$ and $\lambda \geq 2k\sigma\sqrt{\frac{t^2 + 2\log(p+1)}{\min\{n_0, n_1\}}} \geq 4\nu\sigma\sqrt{\frac{t^2 + 2\log(p+1)}{\min\{n_0, n_1\}}}$,

$$\begin{aligned}
\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}} - \beta\|_1 &\leq \frac{9\lambda^2 s}{\phi^2(\beta)} + \frac{9s}{12(1-\frac{\nu}{k})\phi^2(\beta)} + 12\lambda^2(1-\frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n} \\
&\leq \frac{3s}{2\phi^2(\beta)} + \lambda^2\left(\frac{9s}{\phi^2(\beta)} + 12(1-\frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}\right),
\end{aligned}$$

with probability at least $1 - 4\exp(-t^2/2)$.

In the case ii),

$$\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\left(1 - \frac{\nu}{k}\right)\|\hat{\tilde{\beta}} - \beta\|_1 \leq 3\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n},$$

and taking $k \geq 2\nu$ and if $\lambda^2 \geq 1$,

$$\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} + \lambda\|\hat{\tilde{\beta}} - \beta\|_1 \leq 6\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n} \leq 12\lambda^2(1 - \frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}.$$

Thus the result also holds in this case.

□

Now we observe that

$$E\left(\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n}\right) = \int_0^\infty P\left(\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n} > y\right)dy.$$

If we consider the change of variables

$$\begin{aligned} y &= \frac{3s}{2\phi^2(\beta)} + \left(4\nu\sigma\sqrt{\frac{t^2 + 2\log(p+1)}{\min\{n_0, n_1\}}}\right)^2\left(\frac{9s}{\phi^2(\beta)} + 12(1 - \frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}\right) \\ &= \frac{3s}{2\phi^2(\beta)} + \left(16\nu^2\sigma^2\frac{t^2 + 2\log(p+1)}{\min\{n_0, n_1\}}\right)\left(\frac{9s}{\phi^2(\beta)} + 12(1 - \frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}\right) \\ \Rightarrow dy &= \frac{32\nu^2\sigma^2}{\min\{n_0, n_1\}}\left(\frac{9s}{\phi^2(\beta)} + 12(1 - \frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}\right)tdt \end{aligned}$$

then, from previous theorem we have

$$\begin{aligned} E\left(\frac{\|f^* - \tilde{\mathbb{X}}\hat{\tilde{\beta}}\|_2^2}{n}\right) &\leq \frac{32\nu^2\sigma^2}{\min\{n_0, n_1\}}\left(\frac{9s}{\phi^2(\beta)} + 12(1 - \frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}\right)\int_0^\infty 4t\exp(-t^2/2)dt \\ &= \frac{128\nu^2\sigma^2}{\min\{n_0, n_1\}}\left(\frac{9s}{\phi^2(\beta)} + 12(1 - \frac{\nu}{k})\frac{\|\tilde{\mathbb{X}}\beta - f^*\|_2^2}{n}\right). \end{aligned}$$

LASSO estimation, being one of the most common and well-known techniques for assessing the quality of a regression model, makes the computations and results in this section of interest. Yet we believe that the proposed model (4.8.1) could be reconsidered for further analysis of its limitations as well as for improvement as future work of this thesis.

Part II

Asymptotic theory

Chapter 5

A central limit theorem for L_p transportation cost on the real line with application to fairness assessment in machine learning

This chapter corresponds to the publication del Barrio et al. [2019b].

Contents

We provide a Central Limit Theorem for the Monge-Kantorovich distance between two empirical distributions with sizes n and m, $\mathcal{W}_p(P_n, Q_m)$, $p \geq 1$, for observations on the real line. In the case $p > 1$ our assumptions are sharp in terms of moments and smoothness. We prove results dealing with the choice of centering constants. We provide a consistent estimate of the asymptotic variance which enables to build two sample tests and confidence intervals to certify the similarity between two distributions. These are then used to assess a new criterion of data set fairness in classification.

5.1 Introduction

Applications of optimal transportation methods have witnessed a huge development in recent times, in a variety of fields, including machine learning and image processing, among others. The number of significant breakthroughs in the involved numerical procedures can help to understand some of the reasons for this interest. We refer to Chizat et al. [2018] for a more detailed account. In the particular field of statistical inference, despite some early contributions (see, e.g., Munk and Czado [1998], del Barrio et al. [1999a], del Barrio et al. [2005] or Freitag et al. [2007]), progress has been more slow. Among the reasons for this different rythm we can quote the claim from Sommerfeld and Munk [2018] that transportation cost distance 'is an attractive tool for

data analysis but statistical inference is hindered by the lack of distributional limits'. Let us try to give a more complete perspective on this claim.

With inferential goals in mind, the main object of interest is the transportation cost between two sets of random points or between an empirical and a reference measure. In the, by now classical, Kantorovich formulation, for probabilities P and Q on $\mathbb{R}^d$ a transportation plan is a joint probability, say π, on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q. The associated transportation cost is

$$I[\pi] = \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y) d\pi(x,y),$$

where c is some cost function, and the optimal transportation cost is the minimal value of $I[\pi]$ among all choices of transportation plans, π, between P and Q. The problem admits a much more general formulation, but for our present purposes it is enough to know that for the choice $c(x,y) = c_p(x,y) = \|x-y\|^p$, $p \geq 1$, if we denote by $\mathcal{W}_p^p(P,Q)$ the corresponding optimal transportation cost, then $\mathcal{W}_p$ defines a metric in the set $\mathcal{F}_p(\mathbb{R}^d)$ of probabilities on $\mathbb{R}^d$ with finite p-th moment. We refer to Villani [2003] for general background on these facts.

If we observe $X_1, \ldots, X_n$ i.i.d. P, $Y_1, \ldots, Y_m$ i.i.d. Q and write P_n and Q_m for the associated empirical measures, then, assuming that P and Q have finite p-th moment it is well-known that $\mathcal{W}_p^p(P_n, Q) \to \mathcal{W}_p^p(P,Q)$ and $\mathcal{W}_p^p(P_n, Q_m) \to \mathcal{W}_p^p(P,Q)$ almost surely. Enhancing this result with a distributional limit theorem would yield a useful inferential tool in different problems. Early work focused on the case $P = Q$. From an inferential point of view this corresponds to goodness-of-fit problems, with a distributional limit result providing approximate distributions under the null model $P = Q$. In this line we must cite Ajtai et al. [1984] and Talagrand and Yukich [1993] dealing with the case when $P = Q$ is the uniform distribution on the unit hypercube, with later contributions (see Dobrić and Yukich [1995], Fournier and Guillin [2015]) covering an increasingly wider setup. These references dealt with general dimension d, but were not satisfactory for inferential goals, since they only dealt with rates of convergence. Until very recently, distributional limits were only available in the one-dimensional case ($d = 1$). In this case, if $p = 1$ then, under some integrability assumptions $\mathcal{W}_1(P_n, P) = O_P(n^{-1/2})$, with $\sqrt{n}\mathcal{W}_1(P_n, P)$ converging weakly to a non Gaussian limit, see del Barrio et al. [1999b]. If $p > 1$ then it is still possible to get a limiting distribution for $\sqrt{n}\mathcal{W}_p(P_n, P)$, but now integrability assumptions are not enough and the available results require some smoothness conditions on P (and on its density), see del Barrio et al. [1999a] and del Barrio et al. [2005] for the case $p = 2$. Some degree of smoothness (absolute continity of P with positive density on an interval) is, in fact, necessary for boundedness of the sequence $\sqrt{n}E(\mathcal{W}_p(P_n, P))$ if $p > 1$, see Bobkov and Ledoux [2014].

In some statistical applications (in bioequivalence testing, but also in the application to fair learning that we present later) the goal is to provide some statistical certification that the data are not too far from a model, say homogeneity, $P = Q$. Not rejecting the null $H_0 : P = Q$ would be a mere sanity check, but would not provide statistical evidence that the null holds (even approximately). However, this kind of evidence would be granted from rejection of the null $H_0 : \rho(P,Q) \geq \Delta_0$ for some distance ρ. Computation of approximate p-values in this setup would be possible through distributional limit theory for the case $P \neq Q$. Hence, in the case of transportation cost metrics it would be useful to prove a central limit theorem (CLT) for

$$r_n\big(\mathcal{W}_p^p(P_n, Q) - a_n\big) \tag{5.1.1}$$

for some centering a_n and scaling $r_n > 0$ (and similarly for the two-sample case) in the case $P \neq Q$. It would be also useful to guarantee that we can take $a_n = \mathcal{W}_p^p(P,Q)$ as centering constants.

For the metric $\mathcal{W}_2$ (or a trimmed version of it) some limiting results for (6.1.1) were given in Munk and Czado [1998] for one-dimensional data. More recently, Sommerfeld and Munk [2018] handles d-dimensional data and general p, but it is constrained to the case when P and Q are finitely supported (extensions to probabilities with countable support are given in Tameling et al. [2017]). The picture is less complete in the case of continuous distributions. Back to the case $p = 2$, a CLT in general dimension has been provided in del Barrio and Loubes [2019]: if Q has a positive density in the interior of its convex support and P and Q have finite moments of order $4 + \delta$ for some $\delta > 0$, then

$$\sqrt{n}\big(\mathcal{W}_2^2(P_n, Q) - E(\mathcal{W}_2^2(P_n, Q))\big) \to_w N(0, \sigma^2(P, Q)) \tag{5.1.2}$$

for some $\sigma^2(P, Q)$ which is not null if and only if $P \neq Q$. A two-sample version of such results are also given in this work. Note that throughout the paper $\to_w$ denotes weak convergence in probabilities.

In this paper we provide extensions of (6.1.2) to general distances $\mathcal{W}_p$, $p \geq 1$. We cover only the case of one-dimensional data. In turn, from a probabilistic point of view the main contributions of this paper are that *(i)* we prove the analogue of (6.1.2) for general $p > 1$ under sharp moment and smoothness assumptions (Theorem 5.2.1; see also the subsequent comments for discussion about the sharpness of this result) and *(ii)* we show that in the case $p = 1$, when strict convexity of the cost function is lost, non-normal limits can occur, even in the case $P \neq Q$ (Theorem 5.2.4). For the statistical applications that we present, the centering constants in the former CLT's are of crucial importance. We provide general conditions under which $E(\mathcal{W}_p^p(P_n, Q))$ can be replaced by $\mathcal{W}_p^p(P, Q)$ as centering constant in (6.1.2) (Proposition 5.2.6). Combined with a consistent estimator of the asymptotic variance in the CLT's (Proposition 5.2.7), this enables us to define a consistent test

$$H_0 : \mathcal{W}_p(P, Q) \geq \Delta_0 \quad \text{vs} \quad H_a : \mathcal{W}_p(P, Q) < \Delta_0, \tag{5.1.3}$$

that is, a consistent method for gathering statistical evidence to conclude that $\mathcal{W}_p(P, Q) < \Delta_0$.

We would like to note at this point that our approach to prove Theorem 5.2.1 uses the fact that if P and Q are probabilities on the real line with distribution functions (d.f.'s) F and G, respectively, then $\mathcal{W}_p^p(P, Q)$ is simply the L_p-distance between quantile functions, that is,

$$\mathcal{W}_p^p(P, Q) = \int_0^1 |F^{-1} - G^{-1}|^p \tag{5.1.4}$$

(see, e.g., Remark 2.19 in Villani [2003]). For this reason, with some abuse of notation, we will write $\mathcal{W}_p(F, G)$ instead of $\mathcal{W}_p(P, Q)$ in the sequel. We remark, however, that we do not rely on strong approximations for the quantile process (as in Munk and Czado [1998] or del Barrio et al. [1999a], for instance). This kind of approach would require much stronger smoothness assumptions on F. Our technique, in contrast, is much closer to that in del Barrio and Loubes [2019] and (5.1.4) is only used to prove some sharp variance bounds (Propositions 5.5.2 and 5.5.4 and Corollary 5.5.3 in the Appendix).

Currently, the increasingly frequent use of machine learning techniques affects many aspects of our lives. This has yielded to a growing scientific attention to the framework of fair learning. We refer for instance to Romei and Ruggieri [2014a], Pedreschi et al. [2012], Chouldechova [2017] or Friedler et al. [2019] and references therein. In this setting, decisions are made by algorithmic procedures and the main concern is to detect whether decision rules, learnt from variables X, are biased with respect to a subcategory of the population. Formally, the problem consists in forecasting a binary variable $Y \in \{0, 1\}$ using observed covariates $X \in \mathbb{R}^d$, $d \geq 1$, and assuming that the population is divided into two categories that represent a bias, modeled by a protected

variable $S \in \{0,1\}$. A decision rule would be unfair for S when it favours individuals in the main protected group, usually $S = 1$, in the sense that the outcome of the algorithm is not just driven by the values of the covariates X but also by the values of S, leading to treating differently individuals from both groups while they have similar covariates. This discrimination may come from the algorithm or from a biased situation that would have been learnt from the training sample.

In the first situation, many criteria have been given in the recent literature on fair learning to detect whether an algorithm is committing discrimination (see Berk et al. [2017b] or Besse et al. [2018b] for a review). A majority of these definitions consider that the decision should be independent from the protected attribute S. In Berk et al. [2017b], a classifier $g : \mathbb{R}^d \to \{0,1\}$ is said to achieve Statistical Parity, with respect to the joint distribution of (X, S), if

$$\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1). \tag{5.1.5}$$

Therefore, if $\mathcal{L}$ denotes the distribution of a random variable, then Statistical Parity is reached by a classifier g when $\mathcal{L}(g(X) \mid S = 0) = \mathcal{L}(g(X) \mid S = 1)$ and $g(X)$ and S are independent.

Yet, in most real problems the independence described in (5.1.5) is difficult to achieve and, in addition, it refers to a given classification rule when in fact very different classifiers could be trained from the same learning sample. Furthermore, algorithms are usually inaccessible, in the sense that explaining how the classifier is chosen may be seen too intrusive by most companies or it may be simply not possible for many of them to change the way their models are built. To beat these shortchomings, another solution originally proposed in Feldman et al. [2015] and further developed in Gordaliza et al. [2019], tries to look for a condition on the learning sample that ensures that every classifier trained from it is fair. This condition must guarantee that (5.1.5) holds for every classifier $g : \mathbb{R}^d \to \{0,1\}$. If we denote in the following $\mu_s := \mathcal{L}\left(X|S = s\right),\ s \in \{0,1\}$, then this means that μ_0 and μ_1 are equal. But certifying this equality is equivalent to the homogeneity testing problem and, as pointed out before, a goodness-of-fit test does not allow such certification. The most we can aspire to is providing statistical evidence that μ_0 and μ_1 are close. In Section 5.4 we argue in favour of the Wasserstein metrics to measure the distances between the distributions.

As noted above, the CLT's provided in this paper enable to construct a new test to assess the degree of dissimilarity of different distributions, P and Q, using our procedure for testing (5.1.3). In the setup of fair learning, rejecting the null with this test we will be able to statistically certify that the distributions μ_0 and μ_1 are not too different. This will guarantee that the data set is fair, in the sense described above. Additionally, we provide a new way of assessing fairness in machine learning by considering confidence intervals for the degree of dissimilarity between these distributions (with respect to the Wasserstein distance). Also, in the last section, we outline how our fairness assessment procedure can be tuned in order to use it with high-dimensional data.

The remaining sections of this paper are organized as follows. Section 5.2 presents the main results, namely, the CLT's for L_p transportation cost for $p \geq 1$, with additional results dealing with the choice of centering constants and consistent estimation of asymptotic variances. In Section 7.6, we validate the theoretical results supporting the consistency of the variance estimation using simulations for normal and uniform models, which also show that the asymptotically correct rejection rates are achieved and gives insight into the power of the test. Finally, Section 5.4 is devoted to the application of this test to detect unfairness. We first introduce two standard fairness criteria in the fair learning literature called disparate impact (DI) and balanced error rate (BER). Then we present how the testing procedure (5.2.15) and confidence intervals (5.2.14) for the Wasserstein distance would provide a coherent measure of unfairness when dealing with data which have been repaired (i.e modified to promote fairness). Then we

apply our theoretical results to real data, controling the amount of reparation and comparing with DI and BER. These analyses also indicate why our fairness assessment procedure would guarantee more robustness with respect to the outcomes of the classifier than the computation of the DI or the BER. Proofs are gathered in the Appendix.

We end this introduction with some words on notation. F^{-1} denotes the quantile function associated to the distribution function F. sgn denotes the sign function ($\mathrm{sgn}(x) = 1$, $x > 0$, $\mathrm{sgn}(x) = -1$, $x < 0$, $\mathrm{sgn}(0) = 0$) and ℓ denotes the Lebesgue measure on $\mathbb{R}$.

5.2 CLT for L_p transportation cost on the real line

In this section we present the main results in this paper, namely, CLT's for the transportation cost between an empirical measure and a target measure or between two empirical measures. Thus, we will assume that $X_1, \ldots, X_n$ are i.i.d. r.v.'s with law P, $Y_1, \ldots, Y_m$ are r.v.'s with law Q, independent of the X_i's. P and Q will be probabilities on the real line. Hence, they are determined by their distribution functions (d.f.'s), that we will denote by F and G. In fact, it is well known that $\mathcal{W}_p^p(P, Q)$ is simply the L_p-distance bewteen quantile functions, that is,

$$\mathcal{W}_p^p(P, Q) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt,$$

(see, e.g., Remark 2.19 in Villani [2003]). For this reason, with some abuse of notation, we will write $\mathcal{W}_p^p(F, G)$ instead of $\mathcal{W}_p^p(P, Q)$ in the sequel. $\mathcal{F}_q$ will denote the set of probabilities on the real line with finite q-th moment. We will write $F \in \mathcal{F}_q$ with the meaning that the probability with d.f. F belongs to $\mathcal{F}_q$. We will also write F_n (resp. G_m) for the empirical d.f. based on $X_1, \ldots, X_n$ (resp. $Y_1, \ldots, Y_m$).

To present our results, we set $h_p(x) = |x|^p$, $x \in \mathbb{R}$, $p > 1$ and introduce the functions

$$c_p(t; F, G) := \int_{F^{-1}(\frac{1}{2})}^{F^{-1}(t)} h_p'\big(s - G^{-1}(F(s))\big) ds, \quad 0 < t < 1. \tag{5.2.1}$$

We note that $h_p'(x) = p\,\mathrm{sgn}(x)|x|^{p-1}$. Since $F^{-1}(\frac{1}{2}) \le s < F^{-1}(t)$ implies $\frac{1}{2} \le F(s) < t$ while for $F^{-1}(t) \le s < F^{-1}(\frac{1}{2})$ we have $t \le F(s) < \frac{1}{2}$, we see that $c_p(t; F, G)$ is finite for every $t \in (0, 1)$. In fact, we show in Lemma 5.5.1 in the Appendix that, under the assumption $F, G \in \mathcal{F}_{2p}$, $c_p(\cdot; F, G) \in L_2(0, 1)$. This allows us to introduce also

$$\bar{c}_p(t; F, G) := c_p(t; F, G) - \int_0^1 c_p(s; F, G) ds, \quad 0 < t < 1. \tag{5.2.2}$$

We observe that changing $F^{-1}(\frac{1}{2})$ by $F^{-1}(t_0)$ in (5.2.1) would not affect the definition of $\bar{c}_p(\cdot; F, G)$.

It is convenient at this point to introduce the notation

$$\sigma_p^2(F, G) = \int_0^1 \bar{c}_p^2(t; F, G) dt. \tag{5.2.3}$$

Lemma 5.5.1 ensures that $\sigma_p^2(F, G)$ is a finite constant provided F and G have finite moments of order $2p$. Note that $\sigma_p^2(F, G) = 0$ if $F = G$. Otherwise, if $F \neq G$ then $G^{-1} \circ F$, which is the optimal transportation map from F to G, is different from the identity on a set of positive measure and $\sigma_p^2(F, G) > 0$ if F is not a Dirac measure. We remark that $\sigma_p^2(F, G)$ is not, in general, symmetric in F and G.

We are ready now for the main result in this section.

Theorem 5.2.1 (Central Limit Theorem for $\mathcal{W}_p$ with $p > 1$) *Assume that $F, G \in \mathcal{F}_{2p}$ and G^{-1} is continuous on $(0,1)$ and $p > 1$. Then*

(i) If $X_1, \ldots, X_n$ are i.i.d. F and F_n is the empirical d.f. based on the X_i's

$$\sqrt{n}(\mathcal{W}_p^p(F_n, G) - E\mathcal{W}_p^p(F_n, G)) \to_w N(0, \sigma_p^2(F, G)).$$

(ii) If, furthermore, F^{-1} is continuous, $Y_1, \ldots, Y_m$ are i.i.d. G, independent of the X_i's, G_m is the empirical d.f. based on the Y_j's and $\frac{n}{n+m} \to \lambda \in (0,1)$ then

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(F_n, G_m) - E\mathcal{W}_p^p(F_n, G_m)) \to_w N(0, (1-\lambda)\sigma_p^2(F, G) + \lambda\sigma_p^2(G, F)).$$

A proof of this result is given in the Appendix. We would like to make some remarks about Theorem 5.2.1 at this point.

Remark 5.2.2 *There has been a significant interest in empirical transportation costs in recent times in the literature. We should mention at least Fournier and Guillin [2015], giving moment bounds and concentration results for empirical transportation with L_p cost in general dimension, and Bobkov and Ledoux [2014], with a comprehensive discussion of the one-dimensional case. Both papers focus on the case where the law underlying the empirical measure and the target measure are equal (in the setup of Theorem 5.2.1, the case $F = G$). With the more specific goal of CLT's for empirical transportation costs, Sommerfeld and Munk [2018] considers the case when the underlying probabilities are finitely supported, while Tameling et al. [2017] covers probabilities with countable support. The approach in these two cases relies on Hadamard directional differentiability of the dual form of the finite (or countable) linear program associated to optimal transportation. Without the constraint of countable support, del Barrio and Loubes [2019] covers quadratic transportation costs in general dimension.*

There are similarities between the approach in del Barrio and Loubes [2019] and the presentation here, as one can see from a look at our Appendix. We must emphasize some significant differences, however. An obvious one is that here we only deal with one dimensional probabilities. On the other hand, we cover general L_p costs. A more significant difference is that assumptions in Theorem 5.2.1 are sharp. Let us focus on (i) to discuss this point. To make sense of $\mathcal{W}_p^p(F_n, G)$ we must consider G with finite p-th moment. Now, if we want F to satisfy (i) for every G with finite p-th moment, by taking G to be Dirac's measure on 0 we see that

$$\mathcal{W}_p^p(F_n, G) = \frac{1}{n}\sum_{i=1}^{n} |X_i|^p$$

and the condition that F has a finite $2p$-th moment is necessary for the CLT to hold. Then it is easy to check that, $\sigma_p^2(F, G) < \infty$ for all F with finite moment of order $2p$ if and only if G has a finite moment of order $2p$. Thus, the assumption of finite moments of order $2p$ for F and G seems to be a minimal requirement for (i) to hold. We note that for the quadratic cost, $p = 2$, Theorem 4.1 in del Barrio and Loubes [2019] required finite moments of order $4 + \delta$ on P and Q for some $\delta > 0$.

Remark 5.2.3 *Some words on the role of the continuity of G^{-1} in (i) are also in place here. That some sort of regularity of the quantile function is needed for handling the empirical transportation functional in dimension one was observed in Bobkov and Ledoux [2014]. In the case $F = G$, absolute continuity of F^{-1} is a necessary condition for having $E(\mathcal{W}_p(F_n, F)) = O(\frac{1}{\sqrt{n}})$ (Theorem 5.6 in Bobkov and Ledoux [2014]). Continuity of G^{-1} is also related to assumption*

(3) in del Barrio and Loubes [2019]. In fact, that assumption, in the case of one-dimensional probabilities, implies that G is supported in a (possibly unbounded) interval and G^{-1} is differentiable in the interior of that interval. Hence, the regularity assumption in Theorem 5.2.1 is also slightly weaker than that in Theorem 4.1 in del Barrio and Loubes [2019]. We should also note at this point that Theorem 1 in Sommerfeld and Munk [2018], for the case of finitely supported probabilities on the real line corresponds to a case of discontinuity of the quantile functions and this can lead to nonnormal limiting distributions.

When $p = 1$, the function $h_1(x) = |x|$ is no longer differentiable at every point and the function $c_1(t; F, G)$ of (5.2.1) is not well defined in general. It turns out that this can destroy the asymptotic normality of $\mathcal{W}_1(F_n, G)$ in some cases, as we can see in our next result, which is proved in the Appendix. For the sake of brevity we present it for the one sample setup, but it could be adapted to a two sample version.

Theorem 5.2.4 *If F satisfies the integrability assumption*

$$\int_{-\infty}^{\infty} \sqrt{F(t)(1-F(t))}dt < \infty \tag{5.2.4}$$

then

$$\sqrt{n}\big(\mathcal{W}_1(F_n, G) - \mathcal{W}_1(F, G)\big) \to_w \int_{\mathbb{R}} v_F(x)dx,$$

where $v_F(x) = B(F(x))$ if $F(x) > G(x)$, $v_F(x) = -B(F(x))$ if $F(x) < G(x)$, $v_F(x) = |B(F(x))|$ if $F(x) = G(x)$ and B is a Brownian bridge on $[0, 1]$. In particular, if $\ell(F = G) = 0$ then

$$\sqrt{n}\big(\mathcal{W}_1(F_n, G) - \mathcal{W}_1(F, G)\big) \to_w N(0, \sigma_1^2(F, G)),$$

with $\sigma_1^2(F, G) = \int_0^1 c_1^2(t; F, G)dt - \Big(\int_0^1 c_1(t; F, G)dt\Big)^2$ and

$$c_1(t; F, G) := \int_{F^{-1}(\frac{1}{2})}^{F^{-1}(t)} \operatorname{sgn}\big(s - G^{-1}(F(s))\big)ds, \quad 0 < t < 1.$$

The proof of this result is postponed to the Appendix.

Remark 5.2.5 *We remark that under the assumption $\ell(F = G) = 0$ we have $\ell(s : s = G^{-1}(F(s))) = 0$ (see the proof of Theorem 5.2.4 for further details) and we could have written h_1' instead of sgn (with $h_1(x) = |x|$) in the definition of c_1. On the other hand, Theorem 5.2.4 shows that, once the strict convexity of the cost function is lost, nonnormal limits can show up, depending on the size of the set $(F = G)$. In the extreme case $F = G$ we recover that, under* (5.2.4),

$$\sqrt{n}\mathcal{W}_1(F_n, F) \to_w \int_{\mathbb{R}} |B(F(x))|dx.$$

This was part of Theorem 1.1 in del Barrio et al. [1999b]. Condition (5.2.4) *is slightly stronger that the assumption of finite second moments (it holds if F has a moment of order $2 + \delta$). It would be of interest to determine whether, similar to the case $p > 1$, a finite second moment is enough to guarantee weak convergence of $\sqrt{n}(\mathcal{W}_1(F_n, G) - E(\mathcal{W}_1(F_n, G)))$, to a possibly non normal limit. The technique that we have used in this paper does not seem to be give a complete answer to that question.*

We would like to discuss next the role of the centering constants in Theorem 5.2.1. Under more restrictive assumptions there are similar CLT's in which $E\mathcal{W}_p^p(F_n, G)$ is replaced by the simpler constants $\mathcal{W}_p^p(F, G)$ (see, e.g., Theorem 4.3 in del Barrio and Loubes [2019]). In fact, the Kantorovich duality (see, e.g., Villani [2003]) yields that

$$\mathcal{W}_p^p(F, G) = \sup_{(\varphi,\psi)\in\Phi_p} \int \varphi dF + \int \psi dG,$$

where Φ_p is the set of pairs of integrable functions (with respect to F and G, respectively) satisfying $\varphi(x) + \psi(y) \le |x - y|^p$. But this entails $E(\mathcal{W}_p^p(F_n, G)) \ge \sup_{(\varphi,\psi)\in\Phi_p} E\big(\int \varphi dF_n\big) + \int \psi dG = \sup_{(\varphi,\psi)\in\Phi_p} \int \varphi dF + \int \psi dG = \mathcal{W}_p^p(F, G)$. Hence, we can replace the centering constants in Theorem 5.2.1 provided

$$0 \le \sqrt{n}\big(E(\mathcal{W}_p^p(F_n, G)) - \mathcal{W}_p^p(F, G)\big) \to 0. \tag{5.2.5}$$

Finding sharp conditions under which (5.2.5) holds seems to be a delicate issue. We limit ourselves to providing a set of sufficient conditions for it. The case $F = G$ has been considered in Bobkov and Ledoux [2014] and can be handled with simple moment conditions. The general case that we consider here seems to add some smoothness requirements. We limit our discussion to $p \ge 2$. We will assume that F is twice differentiable, with nonvanishing density, f, in the interior of $\mathrm{supp}(F) = \mathrm{cl}\{x : F(x) \notin \{0, 1\}\}$ and satisfies

$$\sup_{t\in(0,1)} \frac{t(1-t)|f'(F^{-1}(t))|}{f^2(F^{-1}(t))} < \infty. \tag{5.2.6}$$

Furthermore, we will assume that

$$\text{for some } s \in (\tfrac{p}{4}, \tfrac{p}{2}), \quad n^s E\mathcal{W}_p^p(F_n, F) \to 0 \quad \text{as } n \to \infty, \tag{5.2.7}$$

$$\frac{1}{\sqrt{n}} \int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{(t(1-t))^{1/2}}{f^2(F^{-1}(t))} dt \to 0, \tag{5.2.8}$$

$$\int_0^1 \int_0^1 \frac{(s \wedge t - st)^2}{f^2(F^{-1}(s))f^2(F^{-1}(t))} dsdt < \infty. \tag{5.2.9}$$

Condition (5.2.6) is a natural condition for approximating the quantile process by a weighted uniform standard process. We refer to del Barrio et al. [2005] for details. The other three conditions are implied by the stronger assumption

$$\int_0^1 \frac{(t(1-t))^{p/2}}{f^p(F^{-1}(t))} dt < \infty. \tag{5.2.10}$$

This condition is, essentially, needed for ensuring that $n^{p/2} E\mathcal{W}_p^p(F_n, F)$ is a bounded sequence, see Bobkov and Ledoux [2014]. We would like to note that, for $p = 2$, (5.2.10) does not hold for Gaussian F, while (5.2.7), (5.2.8) and (5.2.9) do.

With these assumptions we can prove the following.

Proposition 5.2.6 *Assume $p \ge 2$. Under the assumptions of Theorem 5.2.1,*

(i) if F satisfies (5.2.6) *to* (5.2.9) *then* (5.2.5) *holds and, as a consequence,*

$$\sqrt{n}(\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p(F, G)) \to_w N(0, \sigma_p^2(F, G)).$$

(ii) if, furthermore, G satisfies (5.2.6) *to* (5.2.9) *then*

$$\sqrt{\tfrac{nm}{n+m}}(\mathcal{W}_p^p(F_n, G_m) - \mathcal{W}_p^p(F, G)) \to_w N(0, (1-\lambda)\sigma_p^2(F, G) + \lambda\sigma_p^2(G, F)).$$

A similar property has been proved in Berthet et al. [2017] for non-necessarily independent samples. Yet its proof requires stronger assumptions than the one presented here for the independent case. A proof of Proposition 5.2.6 is given in the Appendix. The scheme of proof, in fact, relies on some auxiliary results in del Barrio et al. [2005] that give, through a completely different approach, asymptotic normality of $\sqrt{n}(\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p(F, G))$.

The economy in assumptions that one can gain from dealing with the centering in Theorem 5.2.1 is, in our view, remarkable. Yet providing sharper conditions under which (5.2.5) holds remains an interesting open question.

For the statistical application of Theorem 5.2.1 it is of interest to have a consistent estimator of the asymptotic variances. In the two sample case this can be done as follows. With the standard notation $X_{(j)}$ for the order statistics, define

$$d_{i,n,m}(X, Y) = \sum_{j=2}^{i} \left[\left| X_{(j)} - G_m^{-1}(\tfrac{j-1}{n}) \right|^p - \left| X_{(j-1)} - G_m^{-1}(\tfrac{j-1}{n}) \right|^p \right], \quad i = 2, \ldots, n$$

with $d_{1,n,m}(X, Y) = 0$ and

$$\hat{\sigma}_{1,n,m}^2 = \tfrac{1}{n}\sum_{i=1}^n d_{i,n,m}^2(X, Y) - \left(\tfrac{1}{n}\sum_{i=1}^n d_{i,n,m}(X, Y) \right)^2. \tag{5.2.11}$$

We define $\hat{\sigma}_{2,n,m}^2$ similarly exchanging the roles of the X_i's and the Y_j's. Finally, we set

$$\hat{\sigma}_{n,m}^2 = \tfrac{m}{n+m}\hat{\sigma}_{1,n,m}^2 + \tfrac{n}{n+m}\hat{\sigma}_{2,n,m}^2. \tag{5.2.12}$$

We show next that $\hat{\sigma}_{n,m}^2$ is a consistent estimator of the asymptotic variance in the two sample case in Theorem 5.2.1. A consistent estimator for the asymptotic variance in the one sample case can be obtained similarly. We omit details.

Proposition 5.2.7 *If $F, G \in \mathcal{F}_{2p}$ and F^{-1}, G^{-1} are continuous on $(0, 1)$ then*

$$\hat{\sigma}_{n,m}^2 \to (1-\lambda)\sigma_p^2(F, G) + \lambda\sigma_p^2(G, F)$$

almost surely.

Proof. Simply note that $\hat{\sigma}_{1,n,m}^2 = \int_0^1 \bar{c}_p^2(t; F_n, G_m)dt$ and apply Lemma 5.5.1. □

As a consequence of Propositions 5.2.6 and 5.2.7 we have that if, additionally,

$$F \neq G$$

and F (or G) is not a Dirac measure then

$$\sqrt{\tfrac{nm}{n+m}} \tfrac{(\mathcal{W}_p^p(F_n, G_m) - \mathcal{W}_p^p(F, G))}{\hat{\sigma}_{n,m}} \to_w N(0, 1). \tag{5.2.13}$$

We can use (5.2.13) for statistical applications in several ways. From (5.2.13) we see that

$$\left[\mathcal{W}_p^p(F_n, G_m) \pm \sqrt{\tfrac{n+m}{nm}}\hat{\sigma}_{n,m}\Phi^{-1}(1 - \tfrac{\alpha}{2}) \right] \tag{5.2.14}$$

is a confidence interval for $\mathcal{W}_p^p(F,G)$ with asymptotic confidence level $1-\alpha$. Alternatively, we could consider the testing problem

$$H_0 : \mathcal{W}_p(F,G) \geq \Delta_0, \quad \text{vs} \quad H_a : \mathcal{W}_p(F,G) < \Delta_0, \tag{5.2.15}$$

where Δ_0 is some threshold (to be determined by the practitioner). Rejection of the null assumption in (5.2.15) would yield statistical evidence that the d.f.'s F and G are almost equal. We can handle this problem by rejecting the null if

$$\mathcal{W}_p^p(F_n, G_m) < \Delta_0^p - \sqrt{\tfrac{n+m}{nm}}\hat{\sigma}_{n,m}\Phi^{-1}(1-\alpha). \tag{5.2.16}$$

It follows from (5.2.13) that the test defined by (5.2.16) has asymptotic level α. In the following sections we explore the use of this test for simulations and then for the assessment of fairness of learning algorithms.

5.3 Simulation results

In this section, we first analyze the finite sample performance of the variance estimation given by (5.2.11)-(5.2.12). Then, we check the performance of the testing procedure (5.2.16) for the testing problem (5.2.15) carrying out simulations under both the null and different alternatives. All the simulations are done for different costs $p = 1, 2, 3$.

Consider two independent samples $X_1, \ldots, X_n$ i.i.d. and $Y_1, \ldots, Y_m$ i.i.d. of distributions F and G, respectively, and denote by F_n and G_m the corresponding empirical distribution functions of each sample. We have simulated these samples from the following models.

Example 5.3.1 (Normal model) *Consider $F \sim N(0,1)$ and $G \sim N(\mu, \lambda)$, $(\mu, \lambda) \in \mathbb{R} \times \mathbb{R}^+$. In this location-scale family, we have $G^{-1}(t) = \lambda\Phi^{-1}(t) + \mu$, $t \in (0,1)$, and*

$$\mathcal{W}_p(F,G) = \left(\int_0^1 \left| (1-\lambda)\Phi^{-1}(t) - \mu \right|^p dt \right)^{\frac{1}{p}}, \ p \geq 1. \tag{5.3.1}$$

For $p = 2$, this is simply $\mathcal{W}_2(F,G) = \sqrt{(1-\lambda)^2 + \mu^2}$. Moreover, if $\lambda \neq 1$,

$$c_p(t; F, G) = \frac{1}{1-\lambda}\left[\left| (1-\lambda)\Phi^{-1}(t) - \mu \right|^p - |\mu|^p \right]$$
$$c_p(t; G, F) = \frac{\lambda}{\lambda - 1}\left[\left| (\lambda-1)\Phi^{-1}(t) + \mu \right|^p - |\mu|^p \right] = -\lambda c_p(t; F, G).$$

Note that in the location model, that is when $\lambda = 1$, $\mathcal{W}_p(F,G) = |\mu|$, $p \geq 1$. In this situation,

$$c_p(t; F, G) = -p \cdot sgn(\mu)\, |\mu|^{p-1}\, \Phi^{-1}(t) = -c_p(t; G, F)$$

and $\sigma_p^2(F,G) = \sigma_p^2(G,F) = p^2\mu^{2p-2}$. Hence, we have an exact expression for the asymptotic variance.

Example 5.3.2 (Uniform model) *Consider $F \sim U(0,1)$ and $G \sim U(a,b)$, $a, b \in \mathbb{R}$, $b > a$. In this case, $G^{-1}(t) = a + (b-a)F^{-1}(t) = a + (b-a)t$, $t \in (0,1)$. The Wasserstein distance between the distributions in the location-scale model, that is when the scale parameter is $b-a \neq 1$, is given by*

$$\mathcal{W}_p(F,G) = \left[\frac{1}{(1-(b-a))(p+1)} \left(|1-b|^{p+1} - |a|^{p+1} \right) \right]^{\frac{1}{p}}, \ p \geq 1. \tag{5.3.2}$$

Moreover,

$$c_p(t;F,G) = \frac{1}{1-(b-a)}\left[|(1-(b-a))t-a|^p - |a|^p\right]$$
$$c_p(t;G,F) = -\frac{b-a}{1-(b-a)}\left[|(1-(b-a))t-a|^p - \left|\frac{a}{b-a}\right|^p\right].$$

In the location case, when $b-a=1$, the distance is $\mathcal{W}_p(F,G)=|a|$. We have that

$$c_p(t;F,G) = p\,|a|^{p-1}\,t$$
$$c_p(t;G,F) = p\,|a|^{p-1}\,(a+t),$$

and the asymptotic variances are $\sigma_p^2(F,G)=\sigma_p^2(G,F)=\frac{1}{12}p^2a^{2p-2}$.

First, we illustrate the quality of the variance approximation (5.2.12) for finite n, and assuming equal sizes $n=m$. We have simulated data sets under the normal model in Example 5.3.1 with $\mu=1,\lambda=1$, and the uniform model in Example 5.3.2 with $a=-\frac{1}{2},\ b=\frac{1}{2}$. In Figure 5.1, we can see that the variance estimates are close in the limit to the assymptotic values. Moreover, Table 5.1 shows the MSE$=\frac{1}{N}\sum_{j=1}^{N}\left|\hat{\sigma}_j^2-\sigma^2\right|^2$ of such estimations as a function of the size n of the samples, for large $N=1,000$. We observe that this error tends to 0 as n increases. Convengence seems to be faster for smaller values of p. There is also some indication that convergence is slower for heavier tails. Consequently, choosing the value $p=1$, instead of 2, is more convenient when dealing in practice with observations drawn from distributions with heavy tails.

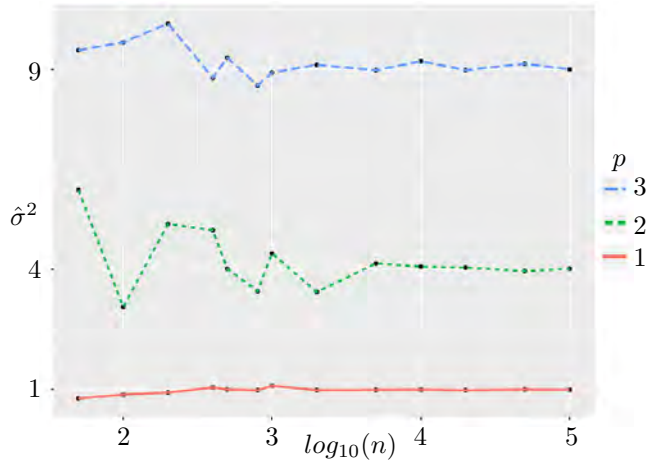
Secondly, to check the performance of the test (5.2.15), we have simulated $1,000$ data sets under the normal and the uniform models for different values of the respective parameters. In Tables 5.2, 5.3 and 5.5 we show the estimated probabilities of rejection for different sample sizes n under the following simulation scenarios:

(i) Normal location model (Table 5.2): $P=N(0,1), Q=N(\mu,1)$, with $\mu=1,0.9,0.7,0.5$, and threshold $\Delta_0=1$

(ii) Normal location-scale model (Table 5.3): $P=N(0,1), Q=N(\mu,\lambda)$, with $(\mu,\lambda)=(1,2)$, $(1,\frac{3}{2}),(0,2),(0,\frac{3}{2})$, and threshold $\Delta_0=\mathcal{W}_p(N(0,1),N(1,2))$

(iii) Uniform model (Table 5.5): $P=U(-\frac{1}{2},\frac{1}{2}), Q=U(a,b)$ with $(a,b)=(-\frac{1}{2},\frac{1}{2}),(-\frac{2}{5},\frac{1}{2})$, $(-\frac{1}{3},\frac{1}{2}),(-\frac{1}{3},\frac{2}{3})$, and threshold $\Delta_0=\mathcal{W}_p(U(0,1),U(-\frac{1}{2},\frac{1}{2}))$.

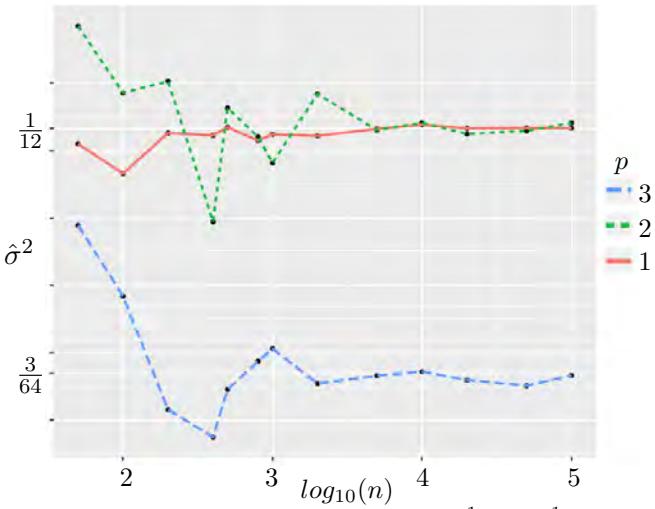
In the third column of each table that corresponds respectively to (i) $\mu=1$, (ii) $(\mu,\lambda)=(1,2)$, (iii) $(a,b)=(-\frac{1}{2},\frac{1}{2})$, such that the null H_0 is true, we observe in all cases a fast convergence of the rejection frequencies to the nominal value $\alpha=0.05$ for every cost p, and from sample sizes not too large. The rest of the columns correspond to situations when the alternative H_a holds. Then, we see that the values of the power are higher as we move away from the boundary of the null hypothesis H_0, without any significant differences in the behavior for different costs p. The Wasserstein distances between the normal distributions in such alternatives are collected in Table 5.4a. When $p=1,3$, the distances (5.3.1) have been numerically computed. In the uniform case, the exact values of the Wasserstein distances (5.3.2) between each distribution $U(a,b)$ and $U(0,1)$ are contained in Table 5.4b.



(a) Normal model with $\mu = 1,\ \lambda = 1$

n	$p=1$	$p=2$	$p=3$
50	0.03076	2.28517	79.70453
100	0.01434	1.25248	36.57057
200	0.00634	0.74908	15.10497
400	0.00290	0.32747	6.15403
500	0.00237	0.21351	5.50914
800	0.00148	0.18638	3.20970
1,000	0.00112	0.13431	2.59728
2,000	0.00054	0.0711	1.41032
5,000	0.00021	0.0304	0.52269
10,000	0.00011	0.0145	0.24127
σ^2	1	4	9

(a) Normal model with $\mu = 1,\ \lambda = 1$



(b) Uniform model with $a = -\frac{1}{2},\ b = \frac{1}{2}$

n	$p=1$	$p=2$	$p=3$
50	8.15575e-05	5.38636e-04	8.20381e-04
100	3.51277e-05	2.91567e-04	3.88538e-04
200	1.55615e-05	1.52519e-04	1.72429e-04
400	7.74579e-06	7.29310e-05	8.51968e-05
500	5.45295e-06	5.60901e-05	7.10645e-05
800	3.98385e-06	3.6331e-05	4.68070e-05
1,000	2.88333e-06	3.12133e-05	3.36132e-05
2,000	1.31779e-06	1.53736e-05	1.59090e-05
5,000	5.63511e-07	6.40119e-06	7.16251e-06
σ^2	1/12	1/12	3/64

(b) Uniform model with $a = -\frac{1}{2},\ b = \frac{1}{2}$

Figure 5.1 – Variance estimates for different sizes n

Table 5.1 – MSE of the variance estimates

5.4 Application to fair learning

Fair learning is devoted to the analysis of biases that may appear when learning automatic decisions, mainly classification rules, from a training sample. This sample may contain some bias against a subpopulation, such that the variable to be predicted is unbalanced between different groups. This bias could have been set intentionally or may reflect the bias present in the use cases. A striking example is provided by Feldman et al. [2015] or Gordaliza et al. [2019], which look at high income prediction from a set of parameters that are influenced by gender. The learning sample includes some numerical attributes together with a high income indicator plus a gender indicator. Imagine that the goal is to train an automatic algorithm from this data to determine whether future employees in a company deserve to be awarded high income. The fact that females in the learning sample are mostly in the low income group can cause that a careless training of an algorithm may associate merit to features which are related to gender, resulting in biased decisions. This gender indicator should not play any role in such forecasts. Thus, it is important to detect such automatic biases in order to prevent their generalization,

p	n	μ=1	μ=0.9	μ=0.7	μ=0.5
1	50	0.062	0.146	0.481	0.825
	100	0.055	0.193	0.698	0.974
	200	0.053	0.275	0.918	1
	400	0.051	0.413	0.995	1
	500	0.051	0.481	0.999	1
	800	0.052	0.64	1	1
	1,000	0.054	0.728	1	1
	2,000	0.047	0.937	1	1
2	50	0.074	0.167	0.513	0.839
	100	0.063	0.198	0.717	0.979
	200	0.059	0.272	0.927	1
	400	0.055	0.422	0.995	1
	500	0.05	0.484	0.999	1
	800	0.053	0.651	1	1
	1,000	0.053	0.736	1	1
	2,000	0.051	0.935	1	1
3	50	0.071	0.154	0.515	0.822
	100	0.0662	0.206	0.715	0.973
	200	0.057	0.266	0.925	1
	400	0.052	0.422	0.992	1
	500	0.057	0.497	0.997	1
	800	0.053	0.652	1	1
	1,000	0.053	0.733	1	1
	2,000	0.051	0.937	1	1

Table 5.2 – Rejection rates in the location normal model with $\Delta_0 = 1$

p	n	$\mu = 1$ $\lambda = 2$	$\mu = 1$ $\lambda = 1.5$	$\mu = 0$ $\lambda = 2$	$\mu = 0$ $\lambda = 1.5$
1	50	0.047	0.165	0.535	0.996
	100	0.045	0.195	0.8	1
	200	0.036	0.323	0.974	1
	400	0.052	0.532	1	1
	500	0.056	0.614	1	1
	800	0.035	0.810	1	1
	1,000	0.045	0.895	1	1
	2,000	0.050	0.994	1	1
2	50	0.078	0.376	0.595	0.998
	100	0.067	0.551	0.823	1
	200	0.062	0.786	0.976	1
	400	0.055	0.969	1	1
	500	0.059	0.985	1	1
	800	0.052	1	1	1
	1,000	0.056	1	1	1
	2,000	0.05	1	1	1
3	50	0.091	0.569	0.571	0.997
	100	0.093	0.762	0.758	1
	200	0.072	0.935	0.939	1
	400	0.06	1	0.996	1
	500	0.064	0.999	0.997	1
	800	0.069	1	1	1
	1,000	0.06	1	1	1
	2,000	0.049	1	1	1

Table 5.3 – Rejection rates in the location-scale normal model when $\Delta_0 = \mathcal{W}_p(N(0,1), N(1,2))$

or even worse, a justification of discriminatory behavior invoking mathematics.

As already mentioned in the introduction, in fair binary classification the data consists in a binary variable $Y \in \{0,1\}$ that we aim to predict using observed covariates $X \in \mathbb{R}^d$, $d \geq 1$, while a protected variable $S \in \{0,1\}$ models the subdivision of the population into two categories. $S = 0$ stands for the minority class. In some approachs to fairness (see, e.g. Feldman et al. [2015] or Chouldechova [2017]) a classifier g is considered to be fair when the conditional distributions $\mathcal{L}(g(X)|S = 0)$ and $\mathcal{L}(g(X)|S = 1)$ are close enough. This is often quantified in the statistical literature using an index called the DI of the classifier g, with respect to (X, S), as follows

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 \mid S = 0)}{\mathbb{P}(g(X) = 1 \mid S = 1)}. \tag{5.4.1}$$

Hence, a classifier g is said not to have DI at level $\tau \in (0, 1]$ if $DI(g, X, S) > \tau$. Note that in some trials, the value $\tau_0 = 0.8$ has been chosen as a legal score to decide whether the discrimination committed by the algorithm is acceptable or not (see e.g. Feldman et al. [2015] or Zafar et al. [2017a]). A related criterion is the BER of g with respect to (X, S)

$$BER(g, X, S) = \frac{\mathbb{P}\left(g(X) = 0 \mid S = 1\right) + \mathbb{P}\left(g(X) = 1 \mid S = 0\right)}{2}. \tag{5.4.2}$$

It describes how the variable S can be learnt by the classification rule g, originally meant to predict the variable Y in the frame where the two classes $S = 0$ and $S = 1$ are balanced in the population. Given $\varepsilon > 0$, S is said to be $\varepsilon-$predictable from X if there exists a classifier $g \in \mathcal{G}$ such that $BER(g, X, S) \leq \varepsilon$. Equivalently, S is not $\varepsilon-$predictable from X if $BER(g, X, S) > \varepsilon$,

p	1	2	3
$\mu = 1$ $\lambda = 2$	1.16664	1.41421	1.61120
$\mu = 1$ $\lambda = 1.5$	1.00849	1.11803	1.20538
$\mu = 0$ $\lambda = 2$	0.79788	1	1.16858
$\mu = 0$ $\lambda = 1.5$	0.39894	0.5	0.58429

(a) Distances $\mathcal{W}_p(N(0,1), N(\mu, \lambda))$

p	1	2	3
$a = -\frac{1}{2}$ $b = \frac{1}{2}$	0.5	0.5	0.5
$a = -\frac{2}{5}$ $b = \frac{1}{2}$	0.45	0.45093	0.45184
$a = -\frac{1}{3}$ $b = \frac{1}{2}$	0.41667	0.41944	0.42215
$a = -\frac{1}{3}$ $b = \frac{2}{3}$	1/3	1/3	1/3

(b) Distances $\mathcal{W}_p(U(0,1), U(a,b))$

Table 5.4 – Wasserstein distances

p	n	$a = -\frac{1}{2}$ $b = \frac{1}{2}$	$a = -\frac{2}{5}$ $b = \frac{1}{2}$	$a = -\frac{1}{3}$ $b = \frac{1}{2}$	$a = -\frac{1}{3}$ $b = \frac{2}{3}$
1	50	0.063	0.249	0.472	0.902
	100	0.047	0.376	0.724	0.998
	200	0.05	0.561	0.93	1
	400	0.05	0.829	0.998	1
	500	0.047	0.885	1	1
	800	0.053	0.978	1	1
	1,000	0.059	0.995	1	1
	2,000	0.055	1	1	1
2	50	0.068	0.268	0.489	0.899
	100	0.072	0.384	0.739	0.989
	200	0.052	0.588	0.915	1
	400	0.047	0.812	0.998	1
	500	0.054	0.886	1	1
	800	0.05	0.973	1	1
	1,000	0.056	0.99	1	1
	2,000	0.062	1	1	1
3	50	0.078	0.278	0.52	0.9
	100	0.07	0.403	0.704	0.992
	200	0.065	0.606	0.9	0.998
	400	0.052	0.843	0.998	1
	500	0.057	0.873	0.998	1
	800	0.061	0.973	1	1
	1,000	0.05	0.996	1	1
	2,000	0.044	1	1	1

Table 5.5 – Frequencies of rejection in the uniform model when $\Delta_0 = \mathcal{W}_p(U(0,1), U(-\frac{1}{2}, \frac{1}{2}))$

for every $g \in \mathcal{G}$. Thus, if $\varepsilon^* := \min_{g \in \mathcal{G}} BER(g, X, S)$ then S is not $\varepsilon-$predictable from X for all $\varepsilon < \varepsilon^*$. From this we can say that ε^* is a global indicator of the fairness of the data. For more details on these criteria and the relationship between them we refer to Gordaliza et al. [2019]. As in the introduction we denote $\mu_s = \mathcal{L}(X \mid S = s)$. Then (see Gordaliza et al. [2019]) the minimum BER over a family of binary classifiers $\mathcal{G}$ can be expressed in terms of the Total Variation distance between the conditioned distributions of the covariates X with respect to the group S to whom they belong

$$\min_{g \in \mathcal{G}} BER(g, X, S) = \frac{1}{2} \left(1 - d_{TV}\left(\mu_0, \mu_1\right)\right). \tag{5.4.3}$$

We see from (5.4.3) that the maximal value of ε^* is $1/2$, which is only achieved in the case of total confusion between the two conditional distributions. This corresponds to complete absence of bias in the training data. Yet, from the statistical point of view we can only certify that the two distributions are close as noted in the introduction. In the assessment of fairness in algorithmic decisions, the conservative choice is to assume the distributions are different, because rejecting the null would provide statistical evidence that μ_0 and μ_1 are close, ensuring some level of fairness. Thus, in view of (5.4.3), one could be tempted to consider the testing problem $H_0 : d_{TV}\left(\mu_0, \mu_1\right) \geq \Delta_0$ vs $H_a : d_{TV}\left(\mu_0, \mu_1\right) < \Delta_0$, for some small $\Delta_0 > 0$. Unfortunately, this is not feasible: there exists no uniformly consistent test for this problem, see Barron [1989]. Consequently, if we want to statistically assess that μ_0 and μ_1 are not too different, we have to choose a better metric. Hence, we propose to use testing procedures for Wasserstein distances

as described in Section 5.2 for the testing problem

$$H_0 : \mathcal{W}_p(\mu_0, \mu_1) \geq \Delta_0 \quad \text{vs} \quad H_a : \mathcal{W}_p(\mu_0, \mu_1) < \Delta_0, \tag{5.4.4}$$

for a small $\Delta_0 > 0$ and $p \geq 1$. Alternatively, we can provide confidence intervals for $\mathcal{W}_p(\mu_0, \mu_1)$ using (5.2.14). We note that, while our testing procedure and confidence intervals have been developed for univariate data, we could extend their applicability by assigning some score $f : \mathbb{R}^d \to \mathbb{R}$ to each observation and then consider the Wasserstein distance between the distributions of such scores conditioned on the two protected groups $\mathcal{L}(f(X) \mid S = s)$, $s \in \{0, 1\}$. In practice, this score will be estimated from the data through some regression model. This may have an impact on the p-values of the corresponding tests or the coverage probability of the confidence intervals but we expect this impact to be limited, particularly for large sample sizes. In our application, we will use a logistic regression for f. Other regression models or machine learning techniques, such as SVM or random forest, could be used for f, depending on the particular problem at hand.

Recently, a number of different techniques have been proposed for transforming the data when lack of fairness is detected, with the goal of removing or reducing the bias (discrimination) in the data. This type of transformation is often called *repairing*. At a population level, these repairing procedures involve modifying the original conditional distributions of the attributes given the protected variable to make them equal (*total repair*) or close enough to each other (*partial repair*), see Feldman et al. [2015], Gordaliza et al. [2019], Hacker and Wiedemann [2017] or Johndrow and Lum [2019].

It is clear that the choice of the distribution to whom the observed μ_s are mapped should convey as much information as possible on the original covariates X. Otherwise, it would hamper the accuracy of the new classification. This constraint led some authors to recommend the use of the so-called Wasserstein barycenter (of order $p = 2$), see Le Gouic and Loubes [2017] and references therein. Statistical justifications for this choice are provided in Gordaliza et al. [2019]. In particular, it is proved that, under some regularity conditions, the excess risk $\mathcal{E}(\tilde{X})$, namely, the difference in minimal classification error without and with the use of the information contained in S, is controlled by a weighted sum of the Wasserstein distances between the original distributions and the distribution chosen for the repair, as follows

$$\mathcal{E}(\tilde{X}) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_{s\sharp} T_s) \right)^{\frac{1}{2}} \tag{5.4.5}$$

for some constant $K > 0$, where T_s is the optimal transport map pushing each μ_s towards the common target. This bound provides some guidelines in the choice of the target distribution since the Wasserstein barycenter of μ_0 and μ_1, with weights π_0 and π_1, minimizes the right hand side of (5.4.5). With this choice the repaired attributes would be $\tilde{X} = T_s(X)$ and we would have $\mathcal{L}(\tilde{X} \mid S = 0) = \mathcal{L}(\tilde{X} \mid S = 1)$.

A particular version of the *partial repair* procedure introduced in Feldman et al. [2015] is called *Geometric Repair*. The authors propose not to move the conditional distributions to the Wasserstein's barycenter but only part towards it on the Wasserstein's geodesic path between μ_0 and μ_1. Let $\lambda \in [0, 1]$ be the amount of repair desired for X. The two *Partially Repaired* conditional distributions for $s \in \{0, 1\}$ are given by the interpolation

$$\mathcal{L}(\tilde{X}_{s,\lambda}) := \mathcal{L}(\lambda T_s(X) + (1 - \lambda)X \mid S = s), \ \lambda \in [0, 1]\,. \tag{5.4.6}$$

We propose to use the confidence intervals (5.2.14) for the Wasserstein distance between the repaired distributions. This will provide a useful insight into the level of reparation needed to

obtain a reasonable degree of fairness.

To illustrate the application of our results to the fair learning problem, we consider the *Adult Income* data set (available at `https://archive.ics.uci.edu/ml/datasets/adult`). It contains $29,825$ instances consisting in the values of 14 attributes, 6 numeric and 8 categorical, and a categorization of each person as having an income of more or less than $\$50,000$ per year. We will just consider the 5 numerical attributes: *Age, Education Level, Capital Gain, Capital Loss* and *Worked hours per week*. A company trying to get an automatic algorithm for deciding whether an employee deserves salary increase could be tempted to train the algorithm on these data using the attributes and the high income indicator. We will train a logistic regression $f(X) = \log(1/(1+\exp(-\beta^T X)),\ \beta \in \mathbb{R}^5$, to base our decisions on this score. We write g for the corresponding logit classifier.

The sensitive attribute to be the potentially protected is the *Gender* (*male* or *female*). In the following, we encode *female* by $S = 0$ and *male* by $S = 1$. The logit classifier presents some bias with respect to the gender in the sense that the learning sample is biased and a female is less likely to be awarded a salary increase compared to a male with similar characteristics. This unfairness is shown in the literature in terms of DI and BER, as discussed in Besse et al. [2018b] and Gordaliza et al. [2019]. Here, we will use the confidence intervals (5.2.14) to assess fairness of the original data set as well as of the repaired versions. In order to improve the interpretability of the comparisons, we have normalized the scores so that the distances $\mathcal{W}_{n,p}^p$ are of similar magnitude as $B\hat{E}R$. For ease of notation, we continue denoting by $f(X)$ the renormalized score. We mention here that all the analyses have been done for the three costs $p = 1, 2, 3$.

Figure 5.2 shows the 95% confidence intervals for $\mathcal{W}_p^p(\mathcal{L}(f(\tilde{X}_{0,\lambda}) \mid S = 0), \mathcal{L}(f(\tilde{X}_{1,\lambda}) \mid S = 1))$, as the amount of repair $\lambda \in [0, 1]$ in (5.4.6) increases. For a better understanding of Figure 5.2, we have included in Figures 5.3a and 5.3b the evolution of the DI and the BER of the logit classifier, respectively, with $\mathcal{W}_{n,p}^p(\mathcal{L}(f(\tilde{X}_{0,\lambda}) \mid S = 0), \mathcal{L}(f(\tilde{X}_{1,\lambda}) \mid S = 1))$, as the amount of repair $\lambda \in [0, 1]$ decreases. In Figure 5.3a, we can see how the Disparate Impact decreases as the Wasserstein distance increases. The standard 0.8 level is attained when the distance is smaller than $0.16, 0.04$ or 0.0125, respectively for $p = 1, 2, 3$, which corresponds to $\lambda = 0.625$. With this level of repair the BER equals 0.485, as we see in Figure 5.3b. Moreover, Figure 5.3b confirms that the closer the distributions are in Wasserstein distance, the more the BER approaches its minimum value 0.5 and, consequently, the less predictable the protected variable is from the outcome of the logit classifier. Finally, we include Figure 5.4 to show the relationship between the DI and BER of the logit g, and Figure 5.5 to see the evolution of the prediction error of the logit classifier as the amount of repair increases.

From this figures we see that, although in general large values of DI and of BER correspond to small values of $\mathcal{W}_p^p$, this last quantity has a different nature since it evaluates the fairness of the whole data set and not simply of a classification rule. We see this in particular in Figures 5.3a and 5.3b. While the confidence interval for DI in the case $\lambda = 0.625$ includes the value $DI = 1$ (perfect fairness for the algorithm) the corresponding confidence intervals for $\mathcal{W}_p^p$ do not include the zero value, indicating that this level of repair is definitely not enough to guarantee fairness of the repaired data.

In this paper, we have restricted ourselves to the computation of the Wasserstein distance on the real line between the distributions of the score given by logistic regression, conditionally given the protected group. Yet, we note that using a multidimensional version of the CLT in del Barrio and Loubes [2019] we could provide a criterion of fairness directly for the observations $X \in \mathbb{R}^d,\ d \geq 1$, by looking at the Wasserstein distance between μ_0 and μ_1. This approach is also supported by result (5.4.5). This will be the subject of a forthcoming work.

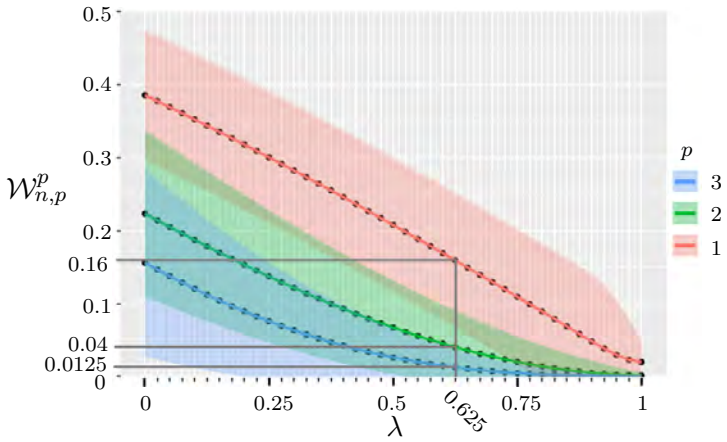


Figure 5.2 – Assymptotic confidence interval for $\mathcal{W}_p^p(\mathcal{L}(f(\tilde{X}_{0,\lambda}) \mid S=0), \mathcal{L}(f(\tilde{X}_{1,\lambda}) \mid S=1))$

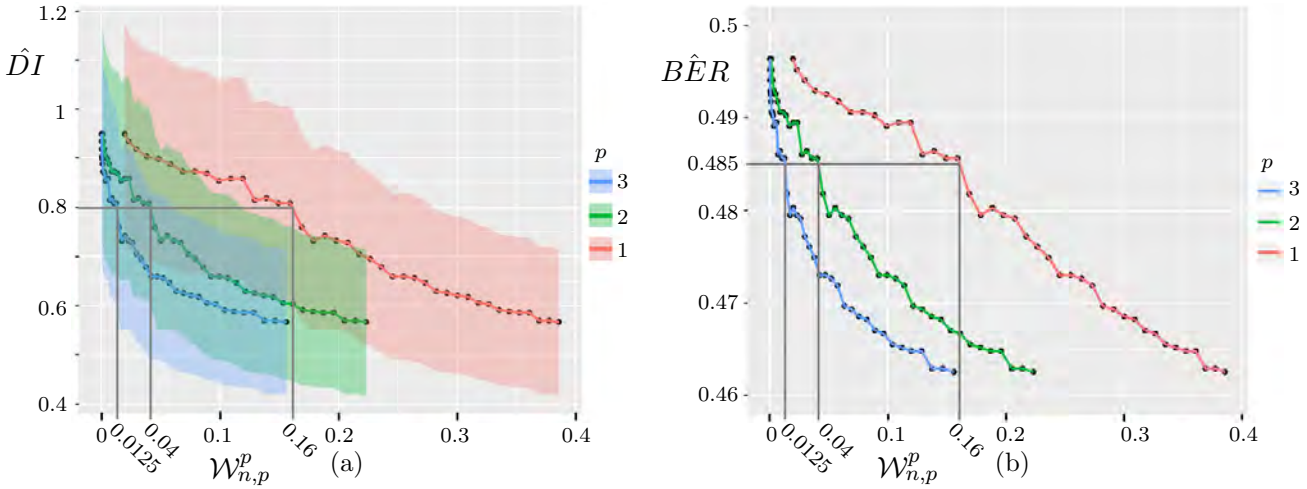


Figure 5.3 – Evolution of (a)$\hat{DI}$ and (b)$B\hat{ER}$ with $\mathcal{W}_{n,p}^p(\mathcal{L}(f(\tilde{X}_{0,\lambda}) \mid S=0), \mathcal{L}(f(\tilde{X}_{1,\lambda}) \mid S=1))$

5.5 Appendix to Chapter 5

In this Appendix we provide the proofs of Theorems 5.2.1 and 5.2.4 and Proposition 5.2.6. For both Theorem 5.2.1 and Proposition 5.2.6 parts i) and ii) can be handled similarly. Hence, for the sake of simplicity we focus on part i). The same techniques yield ii) with little extra effort. Throughout the Appendix we will assume that $U_1, \ldots, U_n$ are i.i.d. r.v.'s uniformly distributed on the interval $(0,1)$. We write A_n for the empirical distribution function on $U_1, \ldots, U_n$ and $\alpha_n(x) = \sqrt{n}(A_n(x) - x)$, $0 \leq x \leq 1$ for the related empirical process. These $U_1, \ldots, U_n$ allow to represent any other i.i.d. sample $X_1, \ldots, X_n$ with d.f. F by taking $X_i = F^{-1}(U_i)$. We use this construction in the sequel without further mention. It will be useful to recall the well known fact (see, e.g., Theorem 6.9 in Villani [2009]) convergence in $\mathcal{W}_p$ metric is equivalent to weak convergence plus convergence of p-th moments. With our notation in terms of d.f.'s this means that $\mathcal{W}_p(F_m, F) \to 0$ if and only if $F_m(x) \to F(x)$ for every continuity point of F and $\int_0^1 |F_m^{-1}(t)|^p dt \to \int_0^1 |F^{-1}(t)|^p dt$ as $m \to \infty$. The convergence condition can be equivalently formulated in terms of quantile functions (see, e.g., Proposition 3.1, p. 112 in Shorack [2000]). Combining this with Vitali's Theorem (see, e.g., Theorem 5.5, p. 55 in Shorack [2000]) we see

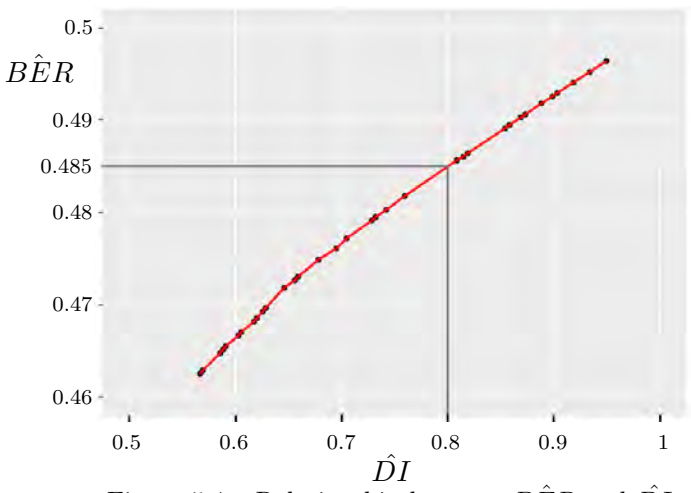


Figure 5.4 – Relationship between $B\hat{E}R$ and $\hat{D}I$

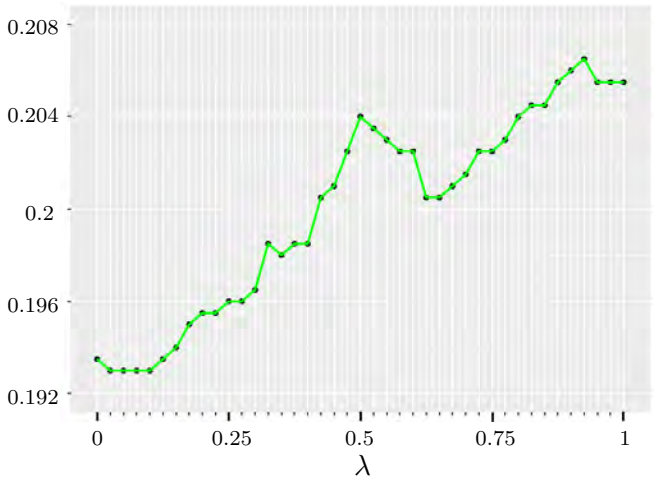


Figure 5.5 – Error in the prediction $g(\tilde{X}_\lambda)$

that

$$\mathcal{W}_p(F_m, F) \to 0 \quad \text{if and only if} \quad \begin{cases} F_m^{-1}(t) \to F^{-1}(t) \text{ for every } t \text{ of continuity of } F^{-1} \\ \text{and } |F_m^{-1}|^p \text{ is uniformly integrable.} \end{cases} \tag{5.5.1}$$

We will make use of (5.5.1) at some points in this Appendix.

Given a distribution function F we write F_n for the empirical distribution function based on the sample $F^{-1}(U_1), \ldots, F^{-1}(U_n)$ and F_n^{-1} for the quantile inverse of F_n. Note that $F_n^{-1}(t) = F^{-1}(A_n^{-1}(t))$. For $p > 1$ we fix a d.f. $G \in \mathcal{F}_{2p}$ and define

$$T_{n,p}(F, G) = \sqrt{n}(\mathcal{W}_p^p(F_n, G) - E(\mathcal{W}_p^p(F_n, G))), \quad F \in \mathcal{F}_{2p}. \tag{5.5.2}$$

Similarly, using the notation in (5.2.1) for c_p and $\bar{c}_p$, we denote

$$T_p(F, G) = \int_0^1 \bar{c}_p(t; F, G) dW(t), \quad F \in \mathcal{F}_{2p},$$

where $\{W(t)\}_{0 \le t \le 1}$ is a standard Brownian motion on $[0, 1]$. Our method of proof of Theorem 5.2.1 is based on a careful analysis of the processes $\{T_{n,p}(F, G)\}_{F \in \mathcal{F}_{2p}}$ and $\{T_p(F, G)\}_{F \in \mathcal{F}_{2p}}$. It follows from Lemma 5.5.1 below and the isometry property of stochastic integrals (see, e.g. chapter 3 in Karatzas and Shreve [1991]) that $T_p(\cdot, G)$ is a centered Gaussian process with covariance function

$$K(F_1, F_2) = \int_0^1 \bar{c}_p(t; F_1, G)\bar{c}_p(t; F_2, G) dt. \tag{5.5.3}$$

In particular, $T_p(F, G)$ is a centered Gaussian r.v. with variance $\sigma_p^2(F, G)$ as in (5.2.3). We observe that our next result shows that $T_p(\cdot, G)$ has continuous trajectories in the sense that $\mathcal{W}_{2p}(F_m, F) \to 0$ implies $E(T_p(F_m, G) - T_p(F, G))^2 \to 0$.

Lemma 5.5.1 *If $F, G \in \mathcal{F}_{2p}$, $p > 1$, then $c_p(\cdot; F, G) \in L_2(0, 1)$ and $\bar{c}_p(\cdot; F, G) \in L_2(0, 1)$. Furthermore, if $F_m, G_m \in \mathcal{F}_{2p}$ satisfy $\mathcal{W}_{2p}(F_m, F) \to 0$, $\mathcal{W}_{2p}(G_m, G) \to 0$ and G^{-1} is continuous on $(0, 1)$ then $\bar{c}_p(\cdot; F_m, G_m) \to \bar{c}_p(\cdot; F, G)$ in $L_2(0, 1)$ as $m \to \infty$.*

Proof. We set $d_p = \max(1, 2^{p-2})$, $p > 1$, and observe that

$$\begin{aligned} |c_p(t; F, G)| &\le p d_p \Big| \int_{F^{-1}(\frac{1}{2})}^{F^{-1}(t)} \big(|s|^{p-1} + |G^{-1}(F(s))|^{p-1}\big) ds \Big| \\ &\le p d_p |F^{-1}(t) - F^{-1}(\tfrac{1}{2})| \Big(|F^{-1}(t)|^{p-1} + |F^{-1}(\tfrac{1}{2})|^{p-1} + |G^{-1}(t)|^{p-1} + |G^{-1}(\tfrac{1}{2})|^{p-1} \Big). \end{aligned} \tag{5.5.4}$$

The first claim follows upon using Hölder's inequality to check that $\int_0^1 |F^{-1}(s)|^2 |G^{-1}(s)|^{2(p-1)} ds < \infty$. For the second we recall from (5.5.1) that $\mathcal{W}_{2p}(F_m, F) \to 0$ implies that $F_m^{-1}(t) \to F^{-1}(t)$ for every t of continuity for F^{-1} (hence, for almost every $t \in (0,1)$) and also that $|F_m^{-1}|^{2p}$ is uniformly integrable (and the same holds for G_m^{-1}, with convergence at every point in $(0,1)$ since G^{-1} is continuous). As noted in Section 5.2, $\bar{c}_p(t; F, G)$ remains unchanged if we replace the integral limit $F^{-1}(\frac{1}{2})$ by a different quantile in the definition of $c_p(t; F, G)$ (as long as we perform the same change in the centering constant). For this reason we can assume without loss of generality that F^{-1} is continuous at $\frac{1}{2}$. Then $|c_p(t; F_m, G)| \to |c_p(t; F, G)|$ at every t of continuity for F^{-1} (pointwise convergence of $h_p'(s - G^{-1}(F_m(s)))$ to $h_p'(s - G^{-1}(F(s)))$ follows from continuity of h_p' and G^{-1}; recall from the discussion after (5.2.1) that for $\varepsilon \in (0, \frac{1}{2})$ and t in $(\varepsilon, 1-\varepsilon)$ all the values of $G^{-1}(F_m(s))$ corresponding to $s \in [\frac{1}{2}, t)$ or $s \in (t, \frac{1}{2}]$ lie in the interval $[G^{-1}(\varepsilon), G^{-1}(1-\varepsilon)]$, which allows us to apply dominated convergence). Now, using the bound (5.5.4) for $c_p(t; F_m, G)$ and uniform integrablility of $|F_m^{-1}|^{2p}$ and $|G_m^{-1}|^{2p}$, we see that the sequence $c_p^2(\cdot; F_m, G_m)$ is uniformly integrable and conclude that $c_p(\cdot; F_m, G) \to c_p(\cdot; F, G)$ and $\bar{c}_p(\cdot; F_m, G_m) \to \bar{c}_p(\cdot; F, G)$ in $L_2(0,1)$.

□

We provide now some empirical counterparts of Lemma 5.5.1. First, a general variance bound for $T_{n,p}(F, G)$ and then, under more restrictive assumptions, an approximate continuity result for the trajectories of $T_{n,p}(\cdot, G)$. The main ingredient in the proof is the Efron-Stein inequality for variances, namely, that if $Z = f(X_1, \ldots, X_n)$ with $X_1, \ldots, X_n$ independent random variables, $(X_1', \ldots, X_n')$ is an independent copy of $(X_1, \ldots, X_n)$ and $Z_i = f(X_1, \ldots, X_i', \ldots, X_n)$ then

$$\mathrm{Var}(Z) \le \sum_{i=1}^n E(Z - Z_i)_+^2.$$

We refer, for instance, to Boucheron et al. [2013] for further details.

Proposition 5.5.2 *If $F, G \in \mathcal{F}_{2p}$, $p > 1$, then there exists a finite constant $C(F, G)$, depending only on F and G such that*

$$\mathrm{Var}\big(T_{n,p}(F, G)\big) \le C(F, G), \quad n \ge 1.$$

A valid choice of the constant is given by $C(F, G) = 8p^2 \max(1, 2^{2(p-1)})(C_1(F) + C_2(F, G))$ with

$$C_1(F) = E\Big(|F^{-1}(U_1) - F^{-1}(U_2)|^2 |F^{-1}(U_1)|^{2(p-1)}\Big)$$

and

$$C_2(F, G) = \Big(E\Big(|F^{-1}(U_1) - F^{-1}(U_2)|^{2p}\Big)\Big)^{1/p} \Big(E\Big(|G^{-1}(U_1)|^{2p}\Big)\Big)^{(p-1)/p}.$$

Proof. We recall that F_n in equation (5.5.2) is the empirical distribution function based on the i.i.d. sample $X_i = F^{-1}(U_i)$, $i = 1, \ldots, n$. We set $Z = \mathcal{W}_p^p(F_n, G)$ and $Z' = \mathcal{W}_p^p(F_n', G)$, where F_n' is the empirical distribution function based on the sample $X_1', X_2, \ldots, X_n$ and $X_1, X_1', X_2 \ldots, X_n$ are i.i.d.. We write $X_{(1)} \le \cdots \le X_{(n)}$ for the ordered sample. Let us assume that F is continuous. Now, $Z = \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} |X_{(i)} - G^{-1}(t)|^p dt = \sum_{i=1}^n \int_{\frac{R_i-1}{n}}^{\frac{R_i}{n}} |X_i - G^{-1}(t)|^p dt$ with R_i denoting the rank of X_i within the sample $X_1, \ldots, X_n$. Continuity of F ensures that a.s. there are no ties and $(R_1, \ldots, R_n)$ is a random permutation of $\{1, \ldots, n\}$. Let us write $(R_1', \ldots, R_n')$ for the ranks in the sample $X_1', X_2, \ldots, X_n$. Now, Z is the minimal value of $E(|U - V|^p | X_1, \ldots, X_n, X_1')$ among

random vectors (U,V) which, conditionally given the X_i's, have marginals F_n and G. This shows that

$$Z \le \sum_{i=1}^{n} \int_{\frac{R_i'-1}{n}}^{\frac{R_i'}{n}} |X_i - G^{-1}(t)|^p dt$$

and, as a consequence,

$$Z - Z' \le \int_{\frac{R_1'-1}{n}}^{\frac{R_1'}{n}} \big[|X_1 - G^{-1}(t)|^p - |X_1' - G^{-1}(t)|^p\big] dt.$$

Using the fact that $||a+h|^p - |a|^p| \le p\,h\,(|a+h|^{p-1} + |a|^{p-1})$ for $a \in \mathbb{R}, h > 0, p > 1$ and writing d_p for the same constants as in the proof of Lemma 5.5.1, we get that

$$\begin{aligned} Z - Z' &\le p|X_1 - X_1'| \int_{\frac{R_1'-1}{n}}^{\frac{R_1'}{n}} \big[|X_1 - G^{-1}(t)|^{p-1} + |X_1' - G^{-1}(t)|^{p-1}\big] dt \\ &\le p d_p |X_1 - X_1'| \Big(2 \int_{\frac{R_1'-1}{n}}^{\frac{R_1'}{n}} |G^{-1}(t)|^{p-1} dt + \frac{|X_1|^{p-1}}{n} + \frac{|X_1'|^{p-1}}{n}\Big). \end{aligned}$$

Hence (observe that R_1 and R_1' are equally distributed),

$$E(Z - Z')_+^2 \le 8p^2 d_p^2 \Big[\tfrac{1}{n^2} E\big(|X_1 - X_1'|^2 |X_1|^{2p-2}\big) + E\big(|X_1 - X_1'| \int_{(R_1-1)/n}^{R_1/n} |G^{-1}(t)|^{p-1} dt\big)^2\Big].$$

Under the assumption $F \in \mathcal{F}_{2p}$, $C_1(F) := E\big(|X_1 - X_1'|^2 |X_1|^{2p-2}\big)$ is finite. To bound the last term we note that,

$$E\Big(|X_1 - X_1'| \textstyle\int_{\frac{R_1-1}{n}}^{\frac{R_1}{n}} |G^{-1}(t)|^{p-1} dt\Big)^2 \le \big(E|X_1 - X_1'|^{2p}\big)^{\frac{1}{p}} \Big(E\Big(\textstyle\int_{\frac{R_1-1}{n}}^{\frac{R_1}{n}} |G^{-1}(t)|^{p-1} dt\Big)^{\frac{2p}{p-1}}\Big)^{\frac{p-1}{p}}.$$

Using again Hölder's inequality we see that

$$\Big(\int_{\frac{j-1}{n}}^{\frac{j}{n}} |G^{-1}(t)|^{p-1} dt\Big)^{\frac{2p}{p-1}} \le n^{-\frac{p+1}{p-1}} \int_{\frac{j-1}{n}}^{\frac{j}{n}} |G^{-1}(t)|^{2p} dt$$

and, therefore,

$$\begin{aligned} E\Big(\int_{\frac{R_1-1}{n}}^{\frac{R_1}{n}} |G^{-1}(t)|^{p-1} dt\Big)^{\frac{2p}{p-1}} &= \frac{1}{n} \sum_{j=1}^{n} \Big(\int_{\frac{j-1}{n}}^{\frac{j}{n}} |G^{-1}(t)|^{p-1} dt\Big)^{\frac{2p}{p-1}} \\ &\le n^{-\frac{2p}{p-1}} \sum_{j=1}^{n} \int_{\frac{j-1}{n}}^{\frac{j}{n}} |G^{-1}(t)|^{2p} dt = \frac{1}{n^{\frac{2p}{p-1}}} \int_0^1 |G^{-1}(t)|^{2p} dt. \end{aligned}$$

As a consequence,

$$E\Big(|X_1 - X_1'| \textstyle\int_{\frac{R_1-1}{n}}^{\frac{R_1}{n}} |G^{-1}(t)|^{p-1} dt\Big)^2 \le \frac{C_2(F,G)}{n^2},$$

with $C_2(F,G) = \big(E|X_1 - X_1'|^{2p}\big)^{\frac{1}{p}} \big(\int_0^1 |G^{-1}(t)|^{2p} dt\big)^{\frac{p-1}{p}} < \infty$. Now the Efron-Stein inequality, and the fact that Z is a symmetric function of $X_1, \ldots, X_n$, which are i.i.d. yields

$$\mathrm{Var}\big(\mathcal{W}_p^p(F_n, G)\big) \le n E(Z - Z')_+^2 \le \frac{C(F,G)}{n}$$

with $C(F,G) = 8p^2 d_p^2(C_1(F) + C_2(F,G))$. This yields the conclusion for continuous F. For general F we take continuous $F_m \in \mathcal{F}_{2p}$ such that $\mathcal{W}_{2p}(F_m, F) \to 0$ as $m \to \infty$ (take as F_m, for instance, the convolution of F with the centered normal distribution with variance $\frac{1}{m}$). A standard uniform integrability argument shows that both $C(F_m, G) \to C(F,G)$ and $\mathrm{Var}\big(T_{n,p}(F_m, G)\big) \to \mathrm{Var}\big(T_{n,p}(F,G)\big)$ as $m \to \infty$ and completes the proof.

□

An interesting consequence of Proposition 5.5.2 is that $T_{n,p}(F,G)$ can be approximated by $T_{n,p}(F_M, G_M)$ with F_M, G_M being bounded support approximations of F and G, respectively. We give details next. We recall that we are using a single uniform sample $U_1, \ldots, U_n$ to generate every empirical d.f., as described at the beginning of this Appendix, and this determines completely the covariance structure of the process $\{T_{n,p}(F,G)\}_{F \in \mathcal{F}_{2p}}$.

Corollary 5.5.3 *Assume $F, G \in \mathcal{F}_{2p}$ and $M > 0$. Consider the distribution function F_M with quantile $F_M^{-1}(t) = \max(\min(F^{-1}(t), M), -M)$. Then there exist constants $C(M,F,G)$ depending only on M, F and G such that*

$$\mathrm{Var}(T_{n,p}(F,G) - T_{n,p}(F_M, G)) \leq C(M,F,G), \quad n \geq 1$$

and $C(M,F,G) \to 0$ as $M \to \infty$. Furthermore, if G_M is the distribution function with quantile $G_M^{-1}(t) = \max(\min(G^{-1}(t), M), -M)$ then for every $\varepsilon > 0$ there exist $M_0 > 0$ and n_0 such that

$$\mathrm{Var}(T_{n,p}(F,G) - T_{n,p}(F_M, G_M)) \leq \varepsilon$$

for each $M \geq M_0$ and $n \geq n_0$.

Proof. We write $\bar{F}_M$ for the distribution function with quantile $\bar{F}_M^{-1}(t) = \min(F^{-1}(t), M)$. We will give a bound for $\mathrm{Var}(T_{n,p}(F,G) - T_{n,p}(\bar{F}_M, G))$, with a similar argument for the left tail completing the proof. Now, observe that

$$\begin{aligned}
\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p((\bar{F}_M)_n, G) &= \int_0^1 |F^{-1}(A_n^{-1}(t)) - G^{-1}(t)|^p dt - \int_0^1 |\bar{F}_M^{-1}(A_n^{-1}(t)) - G^{-1}(t)|^p dt \\
&= \int_{A_n^{-1}(t) > F(M)} |F^{-1}(A_n^{-1}(t)) - G^{-1}(t)|^p dt - \int_{A_n^{-1}(t) > F(M)} |M - G^{-1}(t)|^p dt.
\end{aligned}$$

Note that the last expression does not depend on the values of F^{-1} in the set $\{s \leq F(M)\}$. In particular, if we write $\tilde{F}_M^{-1}(s) = F^{-1}(s)$, if $F^{-1}(s) > M$, $\tilde{F}_M^{-1}(s) = 0$ otherwise, and $\hat{F}_M^{-1}(s) = M$, if $F^{-1}(s) > M$, $\hat{F}_M^{-1}(s) = 0$ otherwise, then $\mathcal{W}_p^p(F_n, G) - \mathcal{W}_p^p((\bar{F}_M)_n, G) = \mathcal{W}_p^p((\tilde{F}_M)_n, G) - \mathcal{W}_p^p((\hat{F}_M)_n, G)$. As a consequence,

$$\mathrm{Var}(T_{n,p}(F,G) - T_{n,p}(\bar{F}_M, G)) \leq 2\mathrm{Var}(T_{n,p}(\tilde{F}_M, G)) + 2\mathrm{Var}(T_{n,p}(\hat{F}_M, G)).$$

It follows from Proposition 5.5.2 that (denoting $a_p = 8p^2 \max(1, 2^{2(p-1)})$

$$\begin{aligned}
\mathrm{Var}(T_{n,p}(\tilde{F}_M, G)) &\leq a_p \Big(\mu_{2p}(\tilde{F}_M)^{\frac{p-1}{p}} + \mu_{2p}(G)^{\frac{p-1}{p}}\Big)\Big(E\Big(|\tilde{F}_M^{-1}(U_1) - \tilde{F}_M^{-1}(U_2)|^{2p}\Big)\Big)^{1/p} \\
&\leq a_p \Big(\mu_{2p}(F)^{\frac{p-1}{p}} + \mu_{2p}(G)^{\frac{p-1}{p}}\Big)\Big(E\Big(|\tilde{F}_M^{-1}(U_1) - \tilde{F}_M^{-1}(U_2)|^{2p}\Big)\Big)^{1/p},
\end{aligned}$$

with $\mu_r(H) = \int_0^1 |H^{-1}(t)|^r dt$. But $\tilde{F}_M^{-1}(U_1) - \tilde{F}_M^{-1}(U_2)$ vanishes if $U_1 \leq F(M)$ and $U_2 \leq F(M)$. Hence,

$$E\Big(|\tilde{F}_M^{-1}(U_1) - \tilde{F}_M^{-1}(U_2)|^{2p}\Big) \leq 2^{2p-1} \int_{(0,1)^2 \setminus (0,F(M))^2} (|F^{-1}(s)|^{2p} + |F^{-1}(t)|^{2p}) ds dt.$$

By dominated convergence the last integral vanishes as $M \to \infty$. To bound $\mathrm{Var}(T_{n,p}(\hat{F}_M, G))$ we simply note that

$$\begin{aligned} E\Big(|\hat{F}_M^{-1}(U_1) - \hat{F}_M^{-1}(U_2)|^{2p}\Big) &= \int_{(0,1)^2} |MI_{\{s \geq F(M)\}} - MI_{\{t \geq F(M)\}}|^{2p} ds dt \\ &\leq 2\int_{F(M)}^1 M^{2p} dt \leq 2\int_{F(M)}^1 |F^{-1}(t)|^{2p} dt \end{aligned}$$

and, again by dominated convergence, the last upper bound vanishes as $M \to \infty$. This proves the first claim and allows to consider only the case of F supported in $[-M, M]$ for the second claim. As before, we show how to deal with the upper tail. Arguing as above, it suffices to bound the variance of $\int_{G(M)}^1 |F_n^{-1}(t) - G^{-1}(t)|^p dt \int_{G(M)}^1 |F_n^{-1}(t) - M|^p dt$. We complete the bound for $Z_M = \int_{G(M)}^1 |F_n^{-1}(t) - G^{-1}(t)|^p dt$, since the other term can be dealt with in a similar way. We consider $X_1' = F^{-1}(U_1')$ with U_1' an independent additional observation with uniform law and argue as in the proof of Proposition 5.5.2. We consider Z_M', the version of Z_M that we obtain replacing X_1 by X_1' in the sample and denote by R_1, R_1' the ranks of X_1 and X_1' in the samples. Now, if $R_1 \leq nG(M)$ and $R_1' \leq nG(M)$ then neither X_1 nor X_1' enter in the expressions that define Z_M and Z_M', respectively, and, consequently, $Z_M - Z_M' = 0$. Also, if $R_1' < R_1$ then $X_1' \leq X_1$ and (recall that $X_1, \ldots, X_n, X_1'$ are upper bounded by M) replacing X_1 by X_1' in the sample can only increase the transportation cost, that is, $Z_M - Z_M' \leq 0$. Hence, if $Z_M - Z_M' > 0$ then $R_1 \leq R_1'$ and $R_1' > nG(M)$. If $R_1 = R_1'$ then $Z_M - Z_M' \leq \int_{\frac{R_1'-1}{n}}^{\frac{R_1'}{n}} \Big||X_1 - G^{-1}(t)|^p - |X_1' - G^{-1}(t)|^p\Big| dt$. If $R_1 < R_1'$ then $X_1 < X_1'$ and from the fact that $a < b < c < d$ implies $(d-b)^p + (c-a)^p \leq (d-a)^p + (b-c)^p$ we can see that $Z_M - Z_M' \leq \int_{\frac{R_1'-1}{n}}^{\frac{R_1'}{n}} \Big||X_1 - G^{-1}(t)|^p - |X_1' - G^{-1}(t)|^p\Big| dt$ as well. Summarizing, we conclude that

$$Z_M - Z_M' \leq \int_{\frac{R_1'-1}{n}}^{\frac{R_1'}{n}} \Big||X_1 - G^{-1}(t)|^p - |X_1' - G^{-1}(t)|^p\Big| dt \, I(R_1' > nG(M)).$$

We can now mimick the proof of Proposition 5.5.2 to see that

$$E(Z_M - Z_M')_+^2 \leq \frac{8p^2 d_p^2}{n^2}\Big(\mu_{2p}(F)^{\frac{p-1}{p}} + \mu_{2p}(G)^{\frac{p-1}{p}}\Big)\Big(E\Big(|X_1 - X_1'|^{2p} I(R_1' > nG(M))\Big)\Big)^{1/p}.$$

Finally, we note that the probability that R_1' exceeds $nG(M)$ is at most $1 - G(M) + \frac{1}{n}$. This completes the proof.

□

When F and G have bounded support and G^{-1} is continuous it is possible to give variance bounds for the increments of $T_{n,p}(\cdot, G)$. In view of Corollary 5.5.3 the assumption of bounded support does not mean a great loss in generality, since slightly worse bounds can be obtained for the general case from this particular one. Please note that the equivalence for the different expressions for $\sigma_p^2(F_1, F_2; G)$ in the next result follows from (5.5.3).

Proposition 5.5.4 *If F_1, F_2 and G are supported in $[-M, M]$ and G^{-1} is continuous then there exists a sequence of constants $R_n(G, p, M)$, which depend on G, p, M and n but not on F_i, $i = 1, 2$ such that $R_n(G, p, M) \to 0$ as $n \to \infty$ and*

$$\mathrm{Var}(T_{n,p}(F_1, G) - T_{n,p}(F_2, G)) \leq 3\sigma_p^2(F_1, F_2; G) + M^2 R_n(G, p, M),$$

with $\sigma_p^2(F_1, F_2; G) = E(T_p(F_1, G) - T_p(F_2, G))^2 = \|\bar{c}_p(\cdot; F_1, G) - \bar{c}_p(\cdot; F_2, G)\|_{L_2(0,1)}^2$.

Proof. We consider first a finitely supported F, concentrated on $x_1 \leq \cdots \leq x_k$ with $F(x_j) = s_j$, $j = 1, \ldots, k$. We have $s_k = 1$ and set, for convenience, $s_0 = 0$. Then $\mathcal{W}_p^p(F,G) = \sum_{j=1}^k \int_{s_{j-1}}^{s_j} |x_j - G^{-1}(t)|^p dt$ and $\mathcal{W}_p^p(F_n,G) = \sum_{j=1}^k \int_{A_n(s_{j-1})}^{A_n(s_j)} |x_j - G^{-1}(t)|^p dt$ (recall the construction for $\alpha_n(x), F_n$ and A_n at the beginning of this Appendix). Hence,

$$\mathcal{W}_p^p(F,G) = \int_0^1 |x_k - G^{-1}(t)|^p dt - \sum_{j=1}^{k-1} \int_0^{s_j} \Big[|x_{j+1} - G^{-1}(t)|^p - |x_j - G^{-1}(t)|^p \Big] dt$$

and similarly for $\mathcal{W}_p^p(F_n,G)$, replacing s_j with $A_n(s_j)$. Writing again $h_p(x) = |x|^p$ we have $|x_{j+1} - G^{-1}(t)|^p - |x_j - G^{-1}(t)|^p = \int_{x_j}^{x_{j+1}} h_p'(s - G^{-1}(t)) ds$ and combining these last two facts, we obtain

$$T_n := \sqrt{n}\Big(\mathcal{W}_p^p(F_n,G) - \mathcal{W}_p^p(F,G) \Big) = -\sqrt{n} \sum_{j=1}^{k-1} \int_{s_j}^{A_n(s_j)} \Big(\int_{x_j}^{x_{j+1}} h_p'(s - G^{-1}(t)) ds \Big) dt. \quad (5.5.5)$$

Next, we define

$$\tilde{T}_n := -\sqrt{n} \sum_{j=1}^{k-1} \int_{s_j}^{A_n(s_j)} \Big(\int_{x_j}^{x_{j+1}} h_p'(s - G^{-1}(s_j)) ds \Big) dt = -\sum_{j=1}^{k-1} \alpha_n(s_j) \int_{x_j}^{x_{j+1}} h_p'(s - G^{-1}(s_j)) ds$$

and observe that

$$|T_n - \tilde{T}_n| \leq \sqrt{n} \sum_{j=1}^{k-1} \Big| \int_{s_j}^{A_n(s_j)} \Big(\int_{x_j}^{x_{j+1}} (h_p'(s - G^{-1}(t)) - h_p'(s - G^{-1}(s_j))) ds \Big) dt \Big|. \quad (5.5.6)$$

For later use, it is convenient to observe that

$$\tilde{T}_n = \sum_{j=1}^{k} (\alpha_n(s_j) - \alpha_n(s_{j-1})) \int_{x_1}^{x_j} h_p'(s - G^{-1}(F(s)))) ds = \int_0^1 \bar{c}_p(s; F, G) d\alpha_n(s)$$

(this is easily checked if one takes into account that F and F^{-1}, and as a consequence $c_p(\cdot; F, G)$, are piecewise constant and also that for any constant k we have $\int_0^1 k d\alpha_n(s) = 0$).

We consider now the continuity moduli

$$w_{G^{-1}}(\delta) = \sup_{x,y \in [0,1], |x-y| \leq \delta} |G^{-1}(x) - G^{-1}(y)|,$$

$$w_{p,M}(\varepsilon) = \sup_{x,y \in [-2M, 2M], |x-y| \leq \varepsilon} |h_p'(x) - h_p'(y)|.$$

The assumptions on G^{-1} imply that it can be extended to a continuous function on $[0,1]$. Hence, it is uniformly continuous and $w_{G^{-1}}(\delta) \to 0$ as $\delta \to 0$. Similarly, $w_{p,M}(\varepsilon) \to 0$ as $\varepsilon \to 0$. Observe now that, for t between s_j and $A_n(s_j)$, $|G^{-1}(t) - G^{-1}(s_j)| \leq w_{G^{-1}}(\|\alpha_n\|_\infty / \sqrt{n})$. Hence,

$$\int_{x_j}^{x_{j+1}} |h_p'(s - G^{-1}(t)) - h_p'(s - G^{-1}(s_j))| ds \leq (x_{j+1} - x_j) w_{p,M}(w_{G^{-1}}(\|\alpha_n\|_\infty / \sqrt{n}))$$

and, therefore, in view of (5.5.6),

$$\begin{aligned} |T_n - \tilde{T}_n| &\leq \sum_{j=1}^{k-1} (x_{j+1} - x_j) w_{p,M}(w_{G^{-1}}(\|\alpha_n\|_\infty / \sqrt{n})) |\alpha_n(s_j)| \\ &\leq \|\alpha_n\|_\infty w_{p,M}(w_{G^{-1}}(\|\alpha_n\|_\infty / \sqrt{n}))(x_k - x_1) \\ &\leq 2M \|\alpha_n\|_\infty w_{p,M}(w_{G^{-1}}(\|\alpha_n\|_\infty / \sqrt{n})). \end{aligned}$$

Hence,

$$E(T_n - \tilde{T}_n)^2 \leq M^2 \tilde{R}_n(G,p,M)$$

with $\tilde{R}_n(G,p,M) = 4E\Big[\|\alpha_n\|_\infty^2 w_{p,M}^2(w_{G^{-1}}(\|\alpha_n\|_\infty/\sqrt{n}))\Big]$. Uniform integrability of $\|\alpha_n\|_\infty^2$ (this follows, for instance, from the well-known Dvoretzky-Kiefer-Wolfowitz inequality, see, e.g., Massart [1990]) and the fact that $w_{p,M}(w_{G^{-1}}(\|\alpha_n\|_\infty/\sqrt{n}))$ is bounded and vanishes in probability ensure that $R_n(G,p,M) \to 0$ as $n \to \infty$.

Let us assume now that F_1 and F_2 are finitely supported as above and write $T_{n,i}$, $\tilde{T}_{n,i}$, $i = 1,2$ for the corresponding versions of T_n and $\tilde{T}_n$, respectively. Observe that there is no loss of generality in assumming that F_1 and F_2 have a common support (observe that (5.5.5) is valid even if $s_{j+1} - s_j = 0$ for some j; we can therefore take the union of the finite supports as the common supporting set). Then

$$\begin{aligned}\mathrm{Var}(T_{n,p}(F_1,G) - T_{n,p}(F_2,G)) &\leq E(T_{n,1} - T_{n,2})^2 \\ &\leq 3E(T_{n,1} - \tilde{T}_{n,1})^2 + 3E(\tilde{T}_{n,1} - \tilde{T}_{n,2})^2 + 3E(T_{n,2} - \tilde{T}_{n,2})^2.\end{aligned}$$

A simple covariance computation using (5.5.5) shows that $E(\tilde{T}_{n,1} - \tilde{T}_{n,2})^2 = \sigma_p^2(F_1,F_2;G)$ and yields the conclusion.

For general F_1 and F_2 we take $F_{i,m}$, $i = 1,2$, $m \geq 1$ with finite support (contained in $[-M,M]$) such that $\mathcal{W}_{2p}(F_{i,m},F_i) \to 0$, $i = 1,2$, and the bound follows by continuity.

□

As a consequence of the variance bounds in Propositions 5.5.2 and 5.5.4 and in Corollary 5.5.3 we can prove now the announced CLT for the empirical transportation cost.

Proof of Theorem 5.2.1. We will prove that

$$\mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F,G)\big), \mathcal{L}\big(T_p(F,G)\big)\big) \to 0$$

As in the proof of Proposition 5.5.3, we assume first that F is concentrated on $x_1 \leq \cdots \leq x_k$ with $F(x_j) = s_j$, $j = 1,\ldots,k$ and F,G supported in $[-M,M]$. Then we have

$$T_n := \sqrt{n}\Big(\mathcal{W}_p^p(F_n,G) - \mathcal{W}_p^p(F,G)\Big) = -\sqrt{n}\sum_{j=1}^{k-1}\int_{s_j}^{A_n(s_j)}\Big(\int_{x_j}^{x_{j+1}} h_p'(s - G^{-1}(t))ds\Big)dt.$$

Continuity of G^{-1} and the multivariate CLT imply that

$$\Big\{\sqrt{n}\int_{s_j}^{A_n(s_j)}\Big(\int_{x_j}^{x_{j+1}} h_p'(s - G^{-1}(t))ds\Big)dt\Big\}_{j=1}^{k-1} \to_w \Big\{B(s_j)\Big(\int_{x_j}^{x_{j+1}} h_p'(s - G^{-1}(s_j))ds\Big)\Big\}_{j=1}^{k-1}$$

as $n \to \infty$, with $B(t)$ a Brownian bridge on $[0,1]$. Hence, using the trivial fact that $\sum_{j=0}^{k-1} B(s_j)c_j = -\sum_{j=0}^{k-1} d_j(B(s_{j+1}) - B(s_j))$ if $d_0 = c_0$ and $d_j = \sum_{l=0}^{j} c_l$, we conclude that

$$T_n \to_w T_p(F,G). \tag{5.5.7}$$

We note that the assumptions on F and G guarantee that

$$\Big|\sqrt{n}\int_{s_j}^{A_n(s_j)}\Big(\int_{x_j}^{x_{j+1}} h_p'(s - G^{-1}(t))ds\Big)dt\Big| \leq K|\alpha_n(s_j)|$$

for some constant K. This shows that T_n^2 is uniformly integrable and, together with (5.5.7) that $\mathcal{W}_2\Big(\mathcal{L}\Big(T_n\Big), \mathcal{L}\Big(T_p(F,G)\Big)\Big) \to 0$. But this, in turn, yields convergence of moments of order 2 or smaller. In particular, we see that $E(T_n) \to E(T_p(F,G)) = 0$, that is

$$\sqrt{n}\big(E(\mathcal{W}_p^p(F_n,G)) - \mathcal{W}_p^p(F,G))\big) \to 0 \tag{5.5.8}$$

as $n \to \infty$. But (5.5.7) and (5.5.8) show that $T_{n,p}(F,G) \to_w T_p(F,G)$ and, again by uniform integrability, that $\mathcal{W}_2\Big(\mathcal{L}\Big(T_{n,p}(F,G)\Big), \mathcal{L}\Big(T_p(F,G)\Big)\Big) \to 0$.

In a second step, we consider F, G supported in $[-M, M]$, with G^{-1} continuous. We consider an approximating sequence F_m with finite support contained in $[-M, M]$ such that $\mathcal{W}_{2p}(F_m, F) \to 0$. Now, for a fixed $\varepsilon > 0$ we can, by Lemma 5.5.1, ensure that $\sigma_p^2(F_m, F, G) \le \varepsilon^2$ for large m. For such m we take n_0 large enough to guarantee that $R_n(G, p, M) \le \varepsilon^2/M^2$ and $\mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F_m,G)\big), \mathcal{L}\big(T_p(F_m,G)\big)\big) \le \varepsilon$ for $n \ge n_0$ (here $R_n(G,p,M)$ is as in Proposition 5.5.4). But then, for $n \ge n_0$,

$$\begin{aligned}\mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F,G)\big), \mathcal{L}\big(T_p(F,G)\big)\big) &\le \mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F,G)\big), \mathcal{L}\big(T_{n,p}(F_m,G)\big)\big) \\ &+ \mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F_m,G)\big), \mathcal{L}\big(T_p(F_m,G)\big)\big) + \mathcal{W}_2\big(\mathcal{L}\big(T_p(F_m,G)\big), \mathcal{L}\big(T_p(F,G)\big)\big) \\ &\le 2\varepsilon + \varepsilon + \varepsilon = 4\varepsilon,\end{aligned}$$

and we conclude that $\mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F,G)\big), \mathcal{L}\big(T_p(F,G)\big)\big) \to 0$ as $n \to \infty$.

Finally, for $F, G \in \mathcal{F}_{2p}$, G^{-1} continuous we use Corollary 5.5.3. Note that G_M^{-1} is also continuous. The already considered cases show that, for fixed M, $\mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F_M,G_M)\big), \mathcal{L}\big(T_p(F_M, G_M)\big)\big) \to 0$ as $n \to \infty$. Obviously, $\mathcal{W}_2\big(\mathcal{L}\big(T_p(F_M,G_M)\big), \mathcal{L}\big(T_p(F,G)\big)\big) \to 0$ as $M \to \infty$. Let us fix $\varepsilon > 0$. We take M_0 and n_0 large enough to ensure that $\mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F,G)\big), \mathcal{L}\big(T_{n,p}(F_M,G_M)\big)\big) \le \epsilon$ if $M \ge M_0$ and $n \ge n_0$ and take $M \ge M_0$ large enough to guarantee $\mathcal{W}_2\big(\mathcal{L}\big(T_p(F_M,G_M)\big), \mathcal{L}\big(T_p(F,G)\big)\big) \le \epsilon$. For this choice of M we take $n_1 \ge n_0$ such that $\mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F_M,G_M)\big), \mathcal{L}\big(T_p(F_M,G_M)\big)\big) \le \varepsilon$ for $n \ge n_1$. But then, arguing as above we see that $\mathcal{W}_2\big(\mathcal{L}\big(T_{n,p}(F,G)\big), \mathcal{L}\big(T_p(F,G)\big)\big) \le 3\varepsilon$ if $n \ge n_1$. This completes the proof.

□

Proof of Proposition 5.2.6. As before, we give a proof for part *(i)*. We will show first that under the given assumptions

$$\sqrt{n}(\mathcal{W}_p^p(F_n,G) - \mathcal{W}_p^p(F,G)) \underset{w}{\to} N(0,\sigma^2) \tag{5.5.9}$$

for some $\sigma^2 \ge 0$. For this goal we note that, by assumption (5.2.7),

$$\sqrt{n}\int_0^1 |F^{-1} - G^{-1}|^{p-2}|F_n^{-1} - F^{-1}|^2 \le \sqrt{n}(\mathcal{W}_p(F,G))^{p-2}(\mathcal{W}_p(F_n,F))^2 = o_P(1).$$

Similarly, we see that

$$\sqrt{n}\int_0^1 |F_n^{-1} - G^{-1}|^{p-2}|F_n^{-1} - F^{-1}|^2 = o_P(1).$$

A Taylor expansion of $h_p(z) = |z|^p$ and the fact that $|x|^{p-2}$ is a convex function imply that

$$\begin{aligned}&\Big||F_n^{-1} - G^{-1}|^p - |F^{-1} - G^{-1}|^p - (F_n^{-1} - F^{-1})h_p'(F^{-1} - G^{-1})\Big| \\ &\qquad \le C(F_n^{-1} - F^{-1})^2\Big(|F^{-1} - G^{-1}|^{p-2} + |F_n^{-1} - G^{-1}|^{p-2}\Big).\end{aligned}$$

This bound and the above estimates yield that

$$\sqrt{n}\big(\mathcal{W}_p^p(F_n,G)-\mathcal{W}_p^p(F,G)-\int_0^1 (F_n^{-1}-F^{-1})h_p'(F^{-1}-G^{-1})\big)=o_P(1).$$

Hence, we focus on the analysis of $\sqrt{n}\int_0^1(F_n^{-1}-F^{-1})h_p'(F^{-1}-G^{-1})$. The moment assumptions on F and G (see, e.g., Lemma 3.3 in Álvarez-Esteban et al. [2011]) allow to replace $\int_0^1$ by $\int_{\frac{1}{n}}^{1-\frac{1}{n}}$ without modyfing the asymptotic behavior of the resulting r.v.. Also, by Lemma 2.3 in del Barrio et al. [2005] and assumptions (5.2.6) and (5.2.8), we can replace $\sqrt{n}(F_n^{-1}-F^{-1})$ in the integral by the weighted uniform quantile process, $u_n/f(F^{-1}(\cdot))$, where $u_n(t)=\sqrt{n}(A_n^{-1}(t)-t)$. Therefore, to prove (5.5.9) it suffices to prove convergence of

$$\int_{1/n}^{1-1/n}\frac{u_n}{f(F^{-1}(\cdot))}h_p'(F^{-1}-G^{-1}).$$

But now, Theorem 4.2 in del Barrio et al. [2005], assumptions (5.2.8) and (5.2.9) and the fact that $h_p'(F^{-1}-G^{-1})$ yield the result.

Now, from (5.5.9) and Theorem 5.2.1 we conclude that $\sqrt{n}(E\mathcal{W}_p^p(F_n,G)-\mathcal{W}_p^p(F,G))$ must be bounded. This in turn yields moment convergence (up to order two; recall the proof of Theorem 5.2.1) of $\sqrt{n}(\mathcal{W}_p^p(F_n,G)-\mathcal{W}_p^p(F,G))$. But since the limiting distribution of $\sqrt{n}(\mathcal{W}_p^p(F_n,G)-\mathcal{W}_p^p(F,G))$ is, as noted above, centered, we must have

$$\sqrt{n}(E\mathcal{W}_p^p(F_n,G)-\mathcal{W}_p^p(F,G))\to 0.$$

This concludes the proof. □

Proof of Theorem 5.2.4. When $p=1$, the identity

$$\mathcal{W}_1(P_n,Q)=\int_{\mathbb{R}}|F_n(x)-G(x)|dx$$

(see, e.g., Villani [2003]) allows to deal with the empirical transportation cost through the consideration of the process

$$\alpha_n^F(x):=\sqrt{n}(F_n(x)-F(x)),\quad x\in\mathbb{R}.$$

Under the assumption

$$\int_{-\infty}^{\infty}\sqrt{F(t)(1-F(t))}dt<\infty$$

we have that α_n^F converges weakly in $L_1(\mathbb{R})$ to B^F, a centered Gaussian process on $\mathbb{R}$ with covariance function

$$\text{Cov }(B^F(x),B^F(y))=F(x\wedge y)-F(x)F(y),$$

see Theorem 2.1 in del Barrio et al. [1999b]. By the Skorohod-Dudley-Wichura Theorem (see, e.g., Theorem 11.7.2 ni Dudley [2002]), we can, therefore, consider versions of α_n^F and B^F such that $\|\alpha_n^F-B^F\|_{L_1}\to 0$ a.s.. Now,

$$\sqrt{n}\big(\mathcal{W}_1(F_n,G)-\mathcal{W}_1(F,G)\big)=\int_{\mathbb{R}}u_n(x)dx,$$

where $u_n(x)=\sqrt{n}\big(|F(x)-G(x)+\alpha_n^F(x)/\sqrt{n}|-|F(x)-G(x)|\big)$. We introduce $v_n(x)=\sqrt{n}\big(|F(x)-G(x)+B^F(x)/\sqrt{n}|-|F(x)-G(x)|\big)$ and $v(x)=B^F(x)$ is $F(x)>G(x)$, $v(x)=$

$-B^F(x)$ if $F(x) < G(x)$ and $v(x) = |B^F(x)|$ if $F(x) = G(x)$. We note that $|u_n(x) - v_n(x)| \leq |\alpha_n^F(x) - B^F(x)|$, which implies that

$$\Big| \int_{\mathbb{R}} u_n(x)dx - \int_{\mathbb{R}} v_n(x)dx \Big| \leq \|\alpha_n^F - B^F\|_{L_1} \to 0 \tag{5.5.10}$$

with probability one.

Now, if $F(x) > G(x)$ then $v_n(x)$ will eventually equal $B^F(x)$, while if $F(x) < G(x)$ then $v_n(x) = -B^F(x)$ for large enough n. Hence, $v_n(x) \to v(x)$ pointwise. On the other hand,

$$|v_n(x)| \leq |B^F(x)|.$$

This shows that we can apply dominated convergence to conclude that

$$\int_{\mathbb{R}} v_n(x)dx \to \int_{\mathbb{R}} v(x)dx. \tag{5.5.11}$$

Combining (5.5.10) and (5.5.11) we see that $\sqrt{n}(\mathcal{W}_1(F_n, G) - \mathcal{W}_1(F, G)) \to \int_{\mathbb{R}} v(x)dx$. To conclude we note that B^F has the same distribution as $B(F(\cdot))$ with B a standard Brownian bridge on $[0, 1]$. Normality and the expression for the variance when $\ell(F = G) = 0$ follow from the fact that, in that case,

$$\int_{\mathbb{R}} v(x)dx = \int_{\mathbb{R}} B(F(x))h(x)dx$$

with $h(x) = I(F(x) > G(x)) - I(F(x) < G(x))$. This last integral is a centered Gaussian r.v. with variance

$$\int_{\mathbb{R}^2} (F(x \wedge y) - F(x)F(y))h(x)h(y)dxdy = \int_0^1 H^2(t)dt - \Big(\int_0^1 H(t)dt \Big)^2,$$

where $H(t) = \int_{F^{-1}(\frac{1}{2})}^{F^{-1}(t)} h(s)ds$ (the last equality follows, from instance, from Proposition 7.4.2, p. 117 in Shorack [2000]). Finally, we note that $F(x) > G(x)$ if and only if $G^{-1}(F(x)) > x$ and also that $x = G^{-1}(F(x))$ if and only if $G(x) \geq F(x)$ and $G(y) < F(x)$ for every $y < x$. But then $G(x) = F(x)$ unless G is not continuous at x. But this can happen at most for a countable collection of x. This means that $I(F(x) > G(x)) = I(G^{-1}(F(x)) > x)$ and, under the assumption $\ell(F = G) = 0$, that $I(F(x) < G(x)) = I(G^{-1}(F(x)) < x)$ for a.e. x. This completes the proof.

□

Chapter 6

Moderate deviations for empirical transportation cost in general dimension

Contents

We provide a Moderate Deviation Principle for the empirical trasportation cost in general dimension. Exploiting the same idea of the linearization approach to obtain the CLT in del Barrio and Loubes [2019], we prove some moment inequalities under more restrictive assumptions. This helps us to analyse the exponential convergence in probability of $\mathcal{W}_2^2(P_n, Q) - \mathbb{E}\mathcal{W}_2^2(P_n, Q)$ towards 0. In the one-dimensional case, we sharpen the moment condition and give a simpler characterization in terms of the speed in the MDP.

6.1 Introduction

Over the last decades, the asymptotic analysis of the optimal transportation cost has been one of the main research topics in probability. Optimal transportation methods have proven to be more and more useful to solve very different real life problems, concerning a wide variety of fields that includes for example imaging sciences (such as color or texture processing), graphics (for shape manipulation) or machine learning (for regression, classification and generative modeling) among others. The significant developments in the numerical procedures that are involved can help to understand some of the reasons for this interest in data analysis. We refer to Chizat et al. [2018] and Peyré et al. [2019] for a more detailed account. In the particular field of statistical inference, despite early contributions in Munk and Czado [1998], del Barrio et al. [1999a], del Barrio et al. [2005] or Freitag et al. [2007], for instance; progress has been slowed by the lack of distributional results. Yet in the latest years this rythm is changing and many generalizations of optimal transport methods have been proposed in relation to approaches originating from statistical inference, such as kernel methods and information theory. We refer to the review Bigot [2019] of the recent contributions in statistics on the use of Wasserstein distances and

tools from optimal transport to analyse datasets whose elements may be modeled as random probability measures, such as multiple histograms or point clouds.

Our main object of interest is the minimal transportation cost between two sets of random points or between an empirical and a reference measure. In the classical Kantorovich formulation the optimal transportation cost between two probabilities P and Q on $\mathbb{R}^d$ is defined as

$$\mathcal{T}(P,Q) := \inf_{\pi\in\Pi(P,Q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\pi(x,y),$$

where $\Pi(P,Q)$ denotes the set of probability measures π over the product space $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q, and c is some cost function. While the problem admits a much more general formulation, for our present purposes it is enough to know that if we denote by $\mathcal{W}_p^p(P,Q)$ the optimal transportation cost corresponding to the choice $c(x,y) = c_p(x,y) = \|x-y\|^p$, $p \geq 1$, then $\mathcal{W}_p$ is the so-called Monge-Kantorovich distance, which defines a metric in the set $\mathcal{F}_p(\mathbb{R}^d)$ of probabilities on $\mathbb{R}^d$ with finite p-th moment. For general background on these facts we refer to Villani [2003].

Recently, much effort has been devoted to the asymptotic analysis of the empirical transportation cost. Precisely, with inference goals in mind, we observe $X_1,\dots,X_n$ i.i.d. with law P, $Y_1,\dots,Y_m$ i.i.d. with law Q, and we write P_n and Q_m for the associated empirical measures. Then, assuming that P and Q have finite p-th moment it is well-known that $\mathcal{W}_p^p(P_n,Q) \to \mathcal{W}_p^p(P,Q)$ and $\mathcal{W}_p^p(P_n,Q_m) \to \mathcal{W}_p^p(P,Q)$ almost surely. Furthermore, it is of great interest to know the rate of such approximation, that is, how far is the empirical transportation cost from its theoretical counterpart. To this task, central limit theorems (CLT), large deviation principles (LDP) and moderate deviation principles (MDP) have been studied for

$$r_n\big(\mathcal{W}_p^p(P_n,Q) - a_n\big),\ p \geq 1, \tag{6.1.1}$$

with some centering a_n and scaling $r_n > 0$ (and similarly for the two-sample case).

The first papers on this topic considered $P = Q$ and $n = m$, meaning that the two random samples come from the same generator. In this case, $\mathcal{W}_p^p(P,Q)$ is exactly zero and the problem is to determined the vanishing rate of the empirical matching transportation cost, namely

$$\mathcal{T}_n^p(P,Q) = \inf_{\sigma\in S_n} \sum_{i=1}^n \left\|X_i - Y_{\sigma(i)}\right\|^p,$$

where S_n denotes the set of permutations of $\{1,\dots,n\}$. Early contributions considered the canonical two-sample matching problem on the plane, that is when $p = 1$, $d = 2$ and $P = Q$ is the uniform distribution on the unit square. In this case, Ajtai et al. [1984] proved that there exists $K > 0$ such that

$$\frac{1}{K}(n\log n)^{1/2} < \mathcal{T}_n^1 < K(n\log n)^{1/2}$$

with probability $1 - o(1)$. Refinements of this result, as well as concentration inequalities, were obtained in Talagrand and Yukich [1993] and Shor [1985], with later extensions in Dobrić and Yukich [1995] and Fournier and Guillin [2015], covering an increasingly wider setup. The connections between the two-sample matching problem and the Monge-Kantorovich problem of optimal transportation of mass were exploited in Ganesh and O'Connell [2007] to obtain moderate and large deviation principles in a fairly general setting. In particular, their main result consist in a MDP over the unit square as well as a LDP over a compact metric space, where the rate function is characterized as a solution to a variational problem. Later Barthe and O'Connell [2009] extended this result to compact support in $\mathbb{R}^d$, and they obtained that the exact moderate deviation rate function on the unit hypercube, is equal to $\frac{(d+2)}{4}x^2$. Their proof

is essentially the same, combining large and moderate deviation results for empirical measures given in Wu [1994] (which relies heavily on earlier work of Ledoux [1992]) with convergence rates for empirical measures in the Monge-Kantorovich distance due to Dudley [1969] and for the unbounded case to Rachev [1991]. A related paper is Gozlan and Léonard [2007] where new transportation cost inequalities are derived by means of elementary large deviation reasonings. Further details on recent developments in the area of transport inequalities could be found in Gozlan and Léonard [2010].

Also a generalization of the moderate deviation principle in Ganesh and O'Connell [2007] to general Polish spaces was proposed in Torrisi [2012]. In this comprehensive work, results on almost sure convergence, large and moderate deviation principles are proved under various assumptions on the reference samples $X_i's, Y_i's$ and the cost function, as well as expressions for the large deviation rate functions in terms of infinite-dimensional variational problems. In some specific situations, more insight into the expressions for the large deviation rate functions is given. In particular, the case when Y_n is supported on some countable subset $\{g_n\}_{n\geq 1} \subset R^d$, which is refered to as the *grid transportation problem*, is deeply studied. The main contributions are: (i) lower bounds for the large deviation rate function of $\mathcal{T}_n^p/n$ for the two-sample matching problem over a compact metric space are provided, as well as a similar result for the grid transportation problem; (ii) the relation between Maurey's $\tau-$property and the large deviation rate function; (iii) for the one-dimensional grid transportation problem over the unit interval $[0, 1]$, the large deviation rate funtion is provided in terms of an optimization problem, which allows its numerical estimation. Moreover, a moderate deviation principle for the optimal transport cost of the grid transportation problem over a compact metric space is proved. Additionally, possible extensions of all of the above to non-compact spaces are briefly discussed. Finally, in the one-dimensional case with compact support, it is shown that the limit distribution for $\mathcal{T}_n^p/\sqrt{n}$ is a random variable whose tail is asymptotically equivalent to the tail of the modulus of a Gaussian random variable. Specifically, a CLT is obtained for the grid transportation problem.

On the other hand and to a lesser extent, the case $P \neq Q$ has also been studied in the literature. We highlight the following contributions to the asymptotic theory in this framework. For one-dimensional data and quadratic cost $p = 2$ some limiting results for (6.1.1) were given in Munk and Czado [1998] for the metric $\mathcal{W}_2$ (or a trimmed version of it). More recently, Sommerfeld and Munk [2018] handles both general cost and dimension for P and Q finitely supported, with later extensions to countable support in Tameling et al. [2017]. In general, few asymptotic results are avalaible in the case of continuous probabilities. Recently, a CLT in general dimension has been provided in del Barrio and Loubes [2019] for quadratic cost $p = 2$: if Q has a positive density in the interior of its convex support and P and Q have finite moments of order $4 + \delta$ for some $\delta > 0$ then

$$\sqrt{n}(\mathcal{W}_2^2(P_n, Q) - \mathbb{E}\mathcal{W}_2^2(P_n, Q)) \to_w N(0, \sigma^2(P, Q)) \tag{6.1.2}$$

for some $\sigma^2(P, Q)$ which is not null if and only if $P \neq Q$. A two-sample version of such results are also given in this work. Note that $\to_w$ denotes weak convergence in probabilities. Extensions of (6.1.2) to general distances $\mathcal{W}_p$, $p \geq 1$, in the one-dimensional case, as well as results dealing with the choice of centering constants were provided in del Barrio et al. [2019b]. The proposed approach in del Barrio and Loubes [2019] is based on the analysis of the optimal transportation potentials, namely, the minimizers in the dual formulation of the optimal transportation problem. Some variance bounds are obtained using the Efron-Stein inequality, that are adapted to prove a linearization result that yields the CLT as a direct consequence.

In this work, we exploit the same idea of the linearization approach in del Barrio and Loubes [2019], under more restrictive assumptions, to prove some moment inequalities that help us to analyse the exponential convergence, in probability, of T_n towards 0. This allows us to obtain a

moderate deviation principle for

$$T_n := \mathcal{W}_2^2(P_n, Q) - \mathbb{E}\mathcal{W}_2^2(P_n, Q).$$

6.2 MDP for empirical transportation cost in general dimension

Over the last years, algorithmic advances have triggered an increasing interest on optimal transport (OT) methods. The popularization of entropic regularization proposed by Cuturi [2013] as a tool of solving large-scale optimal transport problems quickly not only has been shown to yield near-linear-time algorithms for the OT problem [Altschuler et al., 2017], but it also appears to possess useful statistical properties which make it an attractive choice for machine learning applications (see Genevay et al. [2018], Montavon et al. [2016], Rigollet and Weed [2018] Schiebinger et al. [2019]). In particular, with the aim of obtaining results for statistical inference, much effort is being devoted to the asymptotic analysis of the empirical transportation cost in recent researchs. In Mena and Niles-Weed [2019] the authors prove several fundamental statistical bounds for entropic OT with the quadratic Euclidean cost between subgaussian probability measures in arbitrary dimension. Through a new sample complexity result they establish the rate of convergence of entropic OT for empirical measures. Their analysis improves exponentially on the bound of Genevay et al. [2018] and extends their work to unbounded measures. In addition, based on techniques developed by del Barrio and Loubes [2019], they establish a CLT for entropic OT, which was previously only known for finite metric spaces.

Let Z_n be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with values in a topological space $\mathcal{X}$ equipped with the Borel $\sigma-$algebra $\mathcal{B}$. We say that the sequence Z_n satisfies the large deviation principle (LDP) with good rate function I and speed n, if for all $B \in \mathcal{B}$,

$$\begin{aligned} -\inf_{x\in B^o} I(x) &\leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Z_n \in B) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Z_n \in B) \leq -\inf_{x\in \bar{B}} I(x), \end{aligned}$$

where $I \geq 0$, and for $\lambda > 0$ the level sets $\{x : I(x) \leq \lambda\}$ are compact. Let $(a_n)_{n\geq 1}$ be a decreasing, positive sequence such that

$$a_n \to 0 \ \text{ and } \ na_n \to \infty, \text{ as } n \to \infty.$$

We say that the sequence Z_n satisfies the moderate deviation principle (MDP) with good rate function J and speed a_n, if for all $B \in \mathcal{B}$,

$$\begin{aligned} -\inf_{x\in B^o} J(x) &\leq \liminf_{n\to\infty} a_n \log \mathbb{P}(\sqrt{na_n} Z_n \in B) \\ &\leq \limsup_{n\to\infty} a_n \log \mathbb{P}(\sqrt{na_n} Z_n \in B) \leq -\inf_{x\in \bar{B}} J(x), \end{aligned} \tag{6.2.1}$$

where $J \geq 0$, and for $\lambda > 0$ the level sets $\{x : J(x) \leq \lambda\}$ are compact.

The main contribution in our paper is proving a MDP for the empirical OT with quadratic cost in general dimension. Our approach is based on the idea proposed in del Barrio and Loubes [2019] to prove the CLT as in (6.1.2). Let $X_1, \ldots, X_n$ be a sequence of $\mathbb{R}^d-$valued random vectors such that $\log \mathbb{E}\left[e^{\langle \lambda, X_i\rangle}\right] < \infty$ in some ball around the origin, $\mathbb{E}(X_i) = 0$, and the covariance matrix C of X_1 is invertible. Denote the empirical mean $S_n := \frac{1}{n}\sum_{i=1}^n X_i$ and fix $a_n \to 0$ such that $na_n \to \infty$, as $n \to \infty$. From Theorem 3.7.1. in Dembo and Zeitouni

[1998], we have that the sequence S_n satisfies the MDP (6.2.1) with speed a_n and rate function $J(x) = \frac{1}{2}\langle x, C^{-1}x\rangle$.

In the rest of the section, we consider probabilities P and Q in $\mathbb{R}^d$ with positive density in the interior of their convex support, and $X_1, \ldots, X_n$ i.i.d. with law P. Observe that we can write

$$\begin{aligned}\mathcal{W}_2^2(P,Q) &= \min_{\pi\in\Pi(P,Q)} \int \|x-y\|^2 d\pi(x,y) \\ &= \int \|x\|^2 dP + \int \|y\|^2 dQ - 2 \max_{\pi\in\Pi(P,Q)} \int xy d\pi(x,y).\end{aligned}$$

By the Kantorovich Duality (see e.g. Villani [2003]),

$$\max_{\pi\in\Pi(P,Q)} \int xy d\pi(x,y) = \min_{\varphi\in L^1(P), convex} \int \varphi dP + \int \varphi^* dQ,$$

where φ^* denotes the convex conjugate of φ [Rockafellar and Wets, 1998]. Since P has density, the minimizer φ_0 of the right-hand side is the optimal transportation potential from P to Q, up to an additive constant, and $\nabla\varphi_0$ is the optimal transport map (see details in del Barrio and Loubes [2019]). Then, we can write

$$\mathcal{W}_2^2(P,Q) = \int \|x\|^2 dP + \int \|y\|^2 dQ - 2\left(\int \varphi_0 dP + \int \varphi_0^* dQ\right),$$

and similarly for the empirical measure P_n

$$\mathcal{W}_2^2(P_n,Q) = \int \|x\|^2 dP_n + \int \|y\|^2 dQ - 2\left(\int \varphi_n dP_n + \int \varphi_n^* dQ\right),$$

where $\varphi_n,\ n>1$, is the optimal transportation potential from P_n to Q. Thus, we have

$$\begin{aligned}T_n = &\int \|x\|^2 dP_n - \mathbb{E}\|X_1\|^2 \\ &- 2\left[\left(\int \varphi_n dP_n + \int \varphi_n^* dQ\right) - \mathbb{E}\left(\int \varphi_n dP_n + \int \varphi_n^* dQ\right)\right].\end{aligned}$$

Now, we define

$$\begin{aligned}L_n := &\int \|x\|^2 dP_n - \mathbb{E}\|X_1\|^2 \\ &- 2\left[\left(\int \varphi_0 dP_n + \int \varphi_0^* dQ\right) - \mathbb{E}\left(\int \varphi_0 dP_n + \int \varphi_0^* dQ\right)\right] \\ = &\int \|x\|^2 dP_n - 2\int \varphi_0(x) dP_n - \mathbb{E}\|X_1\|^2 + 2\mathbb{E}\varphi_0(X_1),\end{aligned}$$

and we observe that if we denote the random variable $Y_i := \|X_i\|^2 - 2\varphi_0(X_i)$, then we can write

$$L_n = \frac{1}{n}\sum_{i=1}^n Y_i - \mathbb{E}Y_i.$$

Now we define the variance of the variables $Y_i,\ i = 1, \ldots n$,

$$\sigma^2(P,Q) := \mathrm{Var}(\|X_1\|^2 - 2\varphi_0(X_1).$$

Hence, assuming

(I) $\log \mathbb{E}\left[e^{\lambda(Y_i - \mathbb{E}Y_i)}\right] < +\infty,\ \forall \lambda$ with $|\lambda| < \delta$,

(II) $\mathrm{Var}(Y_i) > 0$,

we can deduce from Theorem 3.7.1 in Dembo and Zeitouni [1998] applied to the sequence $Y_1, \ldots, Y_n$, that for any positive sequence $(a_n)_{n>0}$ such that $\lim_{n\to\infty} a_n = 0$ and $\lim_{n\to\infty} na_n = \infty$, we have that for all $t > 0$,

$$-\frac{1}{2}\inf_{x>t}\frac{x^2}{\sigma^2(P,Q)} \leq \liminf_{n\to\infty} a_n \log \mathbb{P}(\sqrt{na_n}L_n > t)$$
$$\leq \limsup_{n\to\infty} a_n \log \mathbb{P}(\sqrt{na_n}L_n > t) \leq -\frac{1}{2}\inf_{x\geq t}\frac{x^2}{\sigma^2(P,Q)}; \tag{6.2.2}$$

and also

$$-\frac{1}{2}\inf_{x<-t}\frac{x^2}{\sigma^2(P,Q)} \leq \liminf_{n\to\infty} a_n \log \mathbb{P}(\sqrt{na_n}L_n < t)$$
$$\leq \limsup_{n\to\infty} a_n \log \mathbb{P}(\sqrt{na_n}L_n < t) \leq -\frac{1}{2}\inf_{x\leq -t}\frac{x^2}{\sigma^2(P,Q)}. \tag{6.2.3}$$

Therefore, L_n satisfies a MDP and our aim is to show the same result in (6.2.2) for T_n. To achieve this property we will obtain an exponential contiguity result for the probability measures $\{\mathcal{L}(L_n)\}_{n\geq 1}$ and $\{\mathcal{L}(T_n)\}_{n\geq 1}$. Precisely, we will show that such measures are exponentially equivalent (see the general Definition 4.2.10 in Dembo and Zeitouni [1998]). By Theorem 4.2.13 in Dembo and Zeitouni [1998] it suffices to prove that for every $\delta > 0$

$$\limsup_{n\to\infty} a_n \log \mathbb{P}(\sqrt{na_n}|T_n - L_n| > \delta) = -\infty. \tag{6.2.4}$$

Consider now $X_1', \ldots, X_n'$ an independent sample copy of $X_1, \ldots, X_n$ and we write $P'_{n,i}$ for the empirical measure on the sample $X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n$. As in the previous work of del Barrio and Loubes [2019], we introduce the rv's

$$R_n := \mathcal{W}_2^2(P_n, Q) - \int \left(\|x\|^2 - 2\varphi_0(x)\right) dP_n(x). \tag{6.2.5}$$

and we notice that $\Delta_n := R_n - \mathbb{E}R_n = L_n - T_n$. For $i = 1, \ldots, n$, let $R_{n,i}$ be computed by replacing in (6.2.5) X_i by X_i' in the sample, that is, replacing P_n by $P'_{n,i}$.

Our approach to obtain the MDP is based on the result in Proposition 6.3.2 postponed to section 6.3, where we prove an upper bound for the moments of the random variable Δ_n. For $r \geq 2$, define the quantities $C_{n,r} := n^r \mathbb{E}(R_n - R_{n,1})_+^r$.

Theorem 6.2.1 *Consider P, Q probabilities with positive density in the interior of their convex support, and $X_1, \ldots, X_n$ i.i.d. with law P. Assume that their support is bounded and fix $\{a_n\}_{n\geq 1}$ such that $a_n \to 0$ and $na_n \to \infty$, as $n \to \infty$. If moreover:*

(A1) P is such that $e^{\lambda\|X_i\|^2} < +\infty,\ \forall \lambda$ with $|\lambda| < \delta$,

(A2) $a_n \log(C_{n,2}) \to -\infty,$ as $n \to \infty$,

then $\{T_n\}_{n\geq 1}$ obeys a MDP with speed $\frac{1}{a_n}$ and good rate function

$$I(x) = \frac{1}{2}\left(\frac{x}{\sigma(P,Q)}\right)^2. \tag{6.2.6}$$

Remark 6.2.2 *With respect to the assumptions in Theorem 6.2.1, we make the following observations:*

1. *By convexity of the exponential,* $e^{\frac{\lambda}{2}Y_i} \leq \frac{1}{2}e^{\lambda\|X_i\|^2} + \frac{1}{2}e^{2\lambda\varphi_0(X_i)}$*, and thus condition* (I) *could be reduced to* $(A1)$*.*

2. *Moreover,* (II) *means that* Y_i *is not constant* $P-a.s.$*, and we note that if* $\varphi_0(x) = \frac{\|x\|^2}{2} - \frac{k}{2}$ $P-a.s.$ *with* $k \in \mathbb{R}$ *constant, then* $\nabla\varphi_0 = Id$*. Hence, condition* (II) *is equivalent to the standard assumption* $P \neq Q$*, which is indeed assumed to obtain the CLT for* T_n *in del Barrio and Loubes [2019].*

Remark 6.2.3 *We note that* $I(x)$ *is in fact a good rate function since for all* $\alpha \in [0, \infty)$ *the level sets* $\{x \in \mathbb{R} | I(x) \leq \alpha\}$ *are compact subsets of* $\mathbb{R}$*.*

Remark 6.2.4 (Condition $(A2)$**)** *We have noticed that verifying condition* $(A2)$ *in general dimension is a complicated task since it amounts to study the rate of vanishing of the quantities* $C_{n,r}$*, for* $r \geq 2$*. This problem has been already tackled in the work of del Barrio and Loubes [2019] where the authors show that if* P *and* Q *have finite moments of order* $4+\delta$*, for some* $\delta > 0$*, then* $C_{n,2} \to 0$ *as* $n \to \infty$*. Refinements on this result remain as future work of this thesis. Yet in the particular one-dimensional setting, this could be sharpened as stated in the following result.*

Corollary 6.2.5 (MDP for probabilities on $\mathbb{R}$**)** *Consider* P *and* Q *probabilities on the real line with respective distribution functions* F *and* G*, and* $X_1, \ldots, X_n$ *i.i.d. with law* P*. Assume moreover that* F *and* G *have positive density and* G^{-1} *is Hölder-continuous. Fix* $\{a_n\}_{n\geq 1}$ *such that* $a_n \to 0$ *and* $na_n \to \infty$*, as* $n \to \infty$*. If moreover:*

(A1) P *is such that* $e^{\lambda|X_i|^2} < +\infty$*,* $\forall\lambda$ *with* $|\lambda| < \delta$*,*

(A2) $a_n \log(n) \to -\infty$*, as* $n \to \infty$*,*

then $\{T_n\}_{n\geq 1}$ *obeys a MDP with speed* $\frac{1}{a_n}$ *and good rate function defined in* (6.2.6)*.*

6.3 Moment bounds for Δ_n

Consider P and Q are probabilities in $\mathbb{R}^d$ with positive density in the interior of their convex support. Let $X_1', \ldots, X_n'$ be an independent sample copy of $X_1, \ldots, X_n$ and we write $P_{n,i}'$ for the empirical measure on the sample $X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n$.

As mentioned above, our approach to obtain the MDP is based on proving an upper bound for the moments of the random variable Δ_n. For $r \geq 2$, recall the quantities $C_{n,r} = n^r\mathbb{E}(R_n - R_{n,1})_+^r$. The following result controls the growth of the moments of Δ_n:

Lemma 6.3.1 *If* P *and* Q *have bounded support, then there exists some constants* $A, \tilde{B} \in \mathbb{R}$ *(depending only on the support of* P, Q*) such that*

$$\mathbb{E}(\Delta_n)^{2q} \leq \tilde{B}A^q \frac{C_{n,r}q!}{n}, \text{ for all } q \geq 1. \tag{6.3.1}$$

Proposition 6.3.2 *If* P *and* Q *have bounded support, then there exists some constants* $A, B \in \mathbb{R}$ *(depending only on the support of* P, Q*) such that for every* $t > 0$*,*

$$\mathbb{P}(\Delta_n \geq t) \leq BC_{n,2}\exp(-\frac{nt^2}{8A}).$$

6.4 Appendix to Chapter 6

Proof of Theorem 6.2.1. From Proposition 6.3.2, we have that for every $t > 0$,

$$\mathbb{P}\left(\sqrt{na_n}|\Delta_n| \geq t\right) = \mathbb{P}\left(|\Delta_n| \geq \frac{t}{\sqrt{na_n}}\right) \leq 2BC_{n,2}\exp(-\frac{t^2}{8Aa_n}).$$

Now, taking logarithms we see that (6.2.4) holds under assumption $(A2)$ since

$$a_n \log \mathbb{P}\left(\sqrt{na_n}|\Delta_n| \geq t\right) \leq a_n \log(2BC_{n,2}) - \frac{t^2}{8A}.$$

Hence, $\{\mathcal{L}(\sqrt{na_n}L_n)\}_{n\geq 1}$ and $\{\mathcal{L}(\sqrt{na_n}T_n)\}_{n\geq 1}$ exponentially equivalent. □

Proof of Lemma 6.3.1. Following Boucheron et al. [2013] we denote the random variables V^+ and V^- by

$$V^+ := \sum_{i=1}^{n} \mathbb{E}[(R_n - R_{n,i})_+^2 | X_1, \ldots, X_n]$$

and

$$V^- := \sum_{i=1}^{n} \mathbb{E}[(R_n - R_{n,i})_-^2 | X_1, \ldots, X_n].$$

For $q \geq 1$, Jensen's inequality for conditional expectations gives

$$\mathbb{E}(V^+)^q = \mathbb{E}\left[\mathbb{E}\left(\sum_{i=1}^{n}(R_n - R_{n,i})_+^2 | X_1, \ldots, X_n\right)\right]^q \leq \mathbb{E}\left(\sum_{i=1}^{n}(R_n - R_{n,i})_+^2\right)^q.$$

Then, we have that $\|V^+\|_q \leq \|\sum_{i=1}^{n}(R_n - R_{n,i})_+^2\|_q$ and by the triangle inequality

$$\|V^+\|_q \leq n\left(\sum_{i=1}^{n}\mathbb{E}(R_n - R_{n,i})_+^{2q}\right)^{\frac{1}{q}} = n\left(\mathbb{E}(R_n - R_{n,1})_+^{2q}\right)^{\frac{1}{q}}.$$

Now we note that

$$\mathcal{W}_2^2(P_n, Q) = \int_{\mathbb{R}^d}(\|x\|^2 - 2\varphi_n(x))dP_n(x) + \int_{\mathbb{R}^d}(\|y\|^2 - 2\varphi_n^*(x))dQ(y)$$

and similarly for $\mathcal{W}_2^2(P_n', Q)$, replacing (φ_n, φ_n^*) with $(\varphi_n', (\varphi_n')^*)$, where φ_n' denotes the optimal transportation potential from P_n' to Q. Also, by optimality,

$$\mathcal{W}_2^2(P_n', Q) \geq \int_{\mathbb{R}^d}(\|x\|^2 - 2\varphi_n(x))dP_n'(x) + \int_{\mathbb{R}^d}(\|y\|^2 - 2\varphi_n^*(x))dQ(y).$$

Hence,

$$\begin{aligned} R_n - R_{n,1} &\leq 2\int_{\mathbb{R}^d}(\varphi_0(x) - \varphi_n(x))dP_n(x) - 2\int_{\mathbb{R}^d}(\varphi_0(x) - \varphi_n(x))dP_n'(x) \\ &= \frac{2}{n}\left[(\varphi_0(X_1) - \varphi_n(X_1)) - \left(\varphi_0(X_1') - \varphi_n(X_1')\right)\right]. \end{aligned}$$

Since the optimal transport maps φ_0, φ_n take values in the bounded support of Q, we deduce that for some positive constant $L \in \mathbb{R}$,

$$(R_n - R_{n,1})_+ \leq \frac{L}{n}$$

Then, we can write

$$\|V^+\|_q \leq n\left(\mathbb{E}(R_n - R_{n,1})_+^{2q}\right)^{\frac{1}{q}} \leq \left(n^q \mathbb{E}(R_n - R_{n,1})_+^r \frac{L^{2q-r}}{n^{2q-r}}\right)^{\frac{1}{q}}$$
$$= \left(C_{n,r}\frac{L^{2q-r}}{n^q}\right)^{\frac{1}{q}}.$$

Similarly, we can bound $\|V^-\|_q$. Then from Theorem 2 in Boucheron et al. [2005] we obtain

$$\|\Delta_n\|_q \leq \|(\Delta_n)_+\|_q + \|(\Delta_n)_-\|_q \leq 2\sqrt{2\kappa q}\sqrt{\frac{C_{n,r}^{\frac{2}{q}} L^{2-\frac{2r}{q}}}{n}}.$$

Hence,

$$\mathbb{E}(\Delta_n)^{2q} \leq 2^{2q}(4\kappa q)^q \frac{C_{n,r}L^{2q-r}}{n} = 16^q \kappa^q q^q \frac{C_{n,r}L^{2q-r}}{n}.$$

Now, by the Stirling's formula there exists $B \in \mathbb{R}$ such that

$$q^q \leq B\frac{q!e^q}{\sqrt{q}} \leq Bq!e^q.$$

We finish the proof taking $\tilde{B} := \frac{B}{L^r}$ and $A := 16\kappa e L^2$. □

Proof of Proposition 6.3.2. Let Δ_n' be an independent copy of $\Delta_n = R_n - \mathbb{E}R_n$. Observe that since Δ_n is centered, $\mathbb{E}e^{-\lambda\Delta_n} \geq 1$, and by symmetry of $\Delta_n' - \Delta_n$ we have

$$\mathbb{E}e^{\lambda\Delta_n} \leq \mathbb{E}e^{\lambda\Delta_n}\mathbb{E}e^{-\lambda\Delta_n} = \mathbb{E}e^{\lambda(\Delta_n - \Delta_n')} = \sum_{q=0}^{\infty}\frac{\lambda^{2q}\mathbb{E}\left[(\Delta_n - \Delta_n')^{2q}\right]}{(2q)!},$$

for every $\lambda \in \mathbb{R}$. Now, by convexity of $x \mapsto x^{2q}$

$$\mathbb{E}\left[(\Delta_n - \Delta_n')^{2q}\right] \leq 2^{2q}\mathbb{E}\left[\Delta_n^{2q}\right].$$

Moreover observe that, for every integer $q \geq 1$,

$$\frac{(2q)!}{q!} = \prod_{j=1}^{q}(q+j) \geq \prod_{j=1}^{q}(2j) = 2^q q!$$

These observations together with previous Lemma 6.3.1 for $r = 2$ imply

$$\mathbb{E}e^{\lambda\Delta_n} \leq \sum_{q=0}^{\infty}\frac{\lambda^{2q}2^{2q}q!BA^qC_{n,2}}{(2q)!n^q} = BC_{n,2}\sum_{q=0}^{\infty}\frac{\lambda^{2q}2^qA^q}{q!n^q} = BC_{n,2}\exp\left(\frac{2\lambda^2 A}{n}\right), \tag{6.4.1}$$

for some constants $A, B \in \mathbb{R}$. Now from Markov's inequality we know that for every $\lambda \geq 0$,

$$\mathbb{P}(\Delta_n \geq t) \leq e^{-\lambda\Delta_n}\mathbb{E}\exp(\lambda\Delta_n).$$

Since this inequality holds for all values of $\lambda \geq 0$, we may choose λ to minimize the upper bound. Let $\psi_{\Delta_n}^*$ be the Cramér transform of Δ_n

$$\psi_{\Delta_n}^* = \sup_{\lambda \geq 0}(\lambda t - \psi_{\Delta_n}(\lambda)), \tag{6.4.2}$$

where $\psi_{\Delta_n}(\lambda) = \log \mathbb{E}e^{\lambda \Delta_n}$, $\lambda \geq 0$, is the logarithm of the moment generating function. Now, consider the function $\phi_n(\lambda) := \log(BC_{n,2}) + \frac{2\lambda^2 A}{n}$. From (6.4.1), we deduce

$$\psi^*_{\Delta_n} \geq \sup_{\lambda \geq 0}(\lambda t - \phi_n(\lambda)) =: \phi^*_n(t). \tag{6.4.3}$$

Furthermore, in del Barrio and Loubes [2019] it is shown that if P and Q have finite moments of order $4+\delta$, for some $\delta > 0$, then $C_{n,2} \to 0$ as $n \to \infty$. This means that for n sufficiently big, $\phi_n(0) \leq 0$ and ϕ^*_n is a nonnegative function. The convexity of the exponential and Jensen's inequality imply $\phi_n(\lambda) \geq \psi(\lambda) \geq \lambda \mathbb{E}\Delta_n$. Hence, for $\lambda < 0$, we have $\lambda t - \phi_n(\lambda) \leq 0$ whenever $t \geq \mathbb{E}\Delta_n$. Consequently, we can extend the supremum over all $\lambda \in \mathbb{R}$ both in (6.4.2) and (6.4.3). For each $n > 0$, the function ϕ_n is continuously differentiable and

$$\frac{d(\lambda t - \phi_n(\lambda))}{d\lambda} = \frac{d}{d\lambda}(\lambda t - \log(BC_{n,2}) - \frac{2\lambda^2 A}{n}) = t - \frac{4\lambda A}{n}.$$

The optimizing value of λ is $\lambda^*_t = \frac{nt}{4A}$ and, as a result,

$$\phi^*_n(t) = \frac{nt^2}{4A} - \log(BC_{n,2}) - \frac{nt^2}{8A} = \frac{nt^2}{8A} - \log(BC_{n,2}).$$

Finally, we conclude since from Chernoff's inequality (see details in Boucheron et al. [2013]) we have that for every $t > 0$,

$$\mathbb{P}\left(\Delta_n \geq t\right) \leq \exp(-\psi^*_{\Delta_n}(t)) \leq \exp(-\phi^*_n(t)) = BC_{n,2}\exp(-\frac{nt^2}{8A}).$$

□

Proof of Corollary 6.2.5. We assume that G^{-1} is Hölder-continuous with exponent $0 < \rho < 1$ and constant $L > 0$. Moreover, we write F_n for the empirical distribution function on $X_1, \ldots, X_n$ and $\alpha^F_n(x) = \sqrt{n}(F_n(x) - F(x))$, $0 \leq x \leq 1$ for the related empirical process. As previously shown in the proof of Lemma 6.3.1, it holds that

$$\begin{aligned} R_n - R_{n,1} &\leq 2\int_{\mathbb{R}^d}(\varphi_0(x) - \varphi_n(x))dP_n(x) - 2\int_{\mathbb{R}^d}(\varphi_0(x) - \varphi_n(x))dP'_n(x) \\ &= \frac{2}{n}\left[(\varphi_0(X_1) - \varphi_n(X_1)) - \left(\varphi_0(X'_1) - \varphi_n(X'_1)\right)\right]. \end{aligned}$$

From this inequality we obtain

$$C_{n,2} = n^2\mathbb{E}(R_n - R_{n,1})^2_+ \leq \frac{8}{n^2}\left[\mathbb{E}\left(\varphi_0(X_1) - \varphi_n(X_1)\right)^2 + \mathbb{E}\left(\varphi_0(X'_1) - \varphi_n(X'_1)\right)^2\right].$$

We fix $y_0 \in \mathbb{R}$ in the interior of the support of G. Since G has density, the optimal transport potential from Q to P_n is

$$\psi_n(y) = \int_{y_0}^{y} F_n^{-1}(G(y))dy,$$

which is a convex and piece-wise linear fucntion, and such that

$$\psi'_n(y) = X_{(i)}, \text{ if } \frac{i-1}{n} < y < \frac{i}{n}.$$

Thus, its convex conjugate $\varphi_n = (\psi_n)^*$ is also a convex and piece-wise linear function with breakpoints $X_{(i)}, i = 1, \ldots, n$, and slope $G^{-1}(\frac{i}{n})$ on each interval $\left(X_{(i)}, X_{(i+1)}\right), i = 1, \ldots, n$.

More precisely, for $X_{(i)} < x < X_{(i+1)}$ we can write

$$\begin{aligned}\varphi_n(x) &= \sum_{j=2}^{i-1} G^{-1}(\frac{j}{n})(X_{(j)} - X_{(j-1)}) + G^{-1}(\frac{i}{n})(x - X_{(i)}) \\ &= \int_{X_{(i)}}^{x} G^{-1}(F_n(s))ds.\end{aligned}$$

Now we observe that for any fixed x_0 in the interior of the support of F, the optimal transport potential from P to Q is the function $\varphi_0(x) = \int_{x_0}^{x} G^{-1}(F(s))ds$, and then we have

$$\varphi_n(x) - \varphi_0(x) = \int_{X_{(i)}}^{x} (G^{-1}(F_n(s)) - G^{-1}(F(s)))ds.$$

From this last expression, we obtain

$$\begin{aligned}|\varphi_n(x) - \varphi_0(x)| &\leq \int_{X_{(i)}}^{x} |G^{-1}(F_n(s)) - G^{-1}(F(s))|ds \\ &\leq \frac{2ML \left\|\alpha_n^F\right\|_\infty^\rho}{n^{\frac{\rho}{2}}}.\end{aligned}$$

Therefore,

$$\mathbb{E}\left(\varphi_0(X_1) - \varphi_n(X_1)\right)^2 \leq C\frac{\mathbb{E}\left(\left\|\alpha_n^F\right\|_\infty^{2\rho}\right)}{n^\rho},$$

and condition $(A2)$ is equivalent to

$$\lim_{n\to\infty} a_n \log\left(C\frac{\mathbb{E}\left(\left\|\alpha_n^F\right\|_\infty^{2\rho}\right)}{n^\rho}\right) = -\infty,$$

which finally means that a_n should be such that $a_n \log(n) \to +\infty$, as $n \to \infty$. □

Chapter 7

Central Limit Theorem and bootstrap procedure for Wasserstein's variations with application to structural relationships between distributions

This chapter corresponds to the publication del Barrio et al. [2019a].

Contents

Wasserstein barycenters and variance-like criterion using Wasserstein distance are used in many problems to analyze the homogeneity of collections of distributions and structural relationships between the observations. We propose the estimation of the quantiles of the empirical process of the Wasserstein's variation using a bootstrap procedure. Then we use these results for statistical inference on a distribution registration model for general deformation functions. The tests are based on the variance of the distributions with respect to their Wasserstein's barycenters for which we prove central limit theorems, including bootstrap versions.

7.1 Introduction

Analyzing the variability of large data sets is a difficult task when the information conveyed by the observations possesses an inner geometry far from the Euclidean one. Indeed, deformations on the data such as translations, scale location models for instance or more general warping procedures prevent the use of the usual methods in statistics. Looking for a way to measure structural relationships between data is of high importance. This kind of issues arises when considering the estimation of probability measures observed with deformations. This situation occurs often in biology, for example when considering gene expression. There has been over the last decade a large amount of work to deal with registrations issues. We refer for instance to Amit et al. [1991], Allassonnière et al. [2007] or Ramsay and Silverman [2005] and references therein. However, when dealing with the registration of warped distributions, the literature is scarce. We mention here the method provided for biological computational issues known as quantile normalization in Bolstad et al. [2003], Gallón et al. [2013] and references therein. Recently, using optimal transport methodologies, comparisons of distributions have been studied using a notion of Fréchet mean for distributions, see for instance in Agueh and Carlier [2011] or a notion of depth as in Chernozhukov et al. [2017].

A natural frame for applications is given by observations drawn from a deformation model in the sense that we observe J independent samples of random variables in $\mathbb{R}$, with sample j following distribution μ_j, such that

$$X_{i,j} = g_j(\varepsilon_{i,j}), \quad j = 1, \dots, J, \quad i = 1 \dots, n,$$

where $(\varepsilon_{i,j})$ are i.i.d. random variables with unknown distribution μ. The functions g_j belong to a class $\mathcal{G}$ of deformation functions, which models how the distributions μ_j's can be warped one to another by functions in the chosen class. This model is the natural extension of the functional deformation models studied in the statistical literature for which estimation procedures are provided in Gamboa et al. [2007] while testing issues are tackled in Collier and Dalalyan [2015]. In the setup of warped distributions a main goal is the estimation of the warping functions, possibly as a first step towards registration or alignment of the (estimated) distributions. Of course, without some constraints on the class $\mathcal{G}$ the deformation model is meaningless (we can, for instance, obtain any distribution on $\mathbb{R}^d$ as a warped version of a fixed probability having a density if we take the optimal transportation map as the warping function; see Villani [2009]) and one has to consider smaller classes of deformation functions to perform a reasonable registration. In the case of parametric classes estimation of the warping functions is studied in Agulló-Antolín et al. [2015]. However, estimation/registration procedures may lead to inconsistent conclusions if the chosen deformation class $\mathcal{G}$ is too small. It is, therefore, important to be able to assess fit to the deformation model given by a particular choice of $\mathcal{G}$ and this is the main goal of this paper. We note that within this framework, statistical inference on deformation models for distributions has been studied first in Freitag and Munk [2005]. Here we provide a different approach which allows to deal with more general deformation classes.

The pioneer works Czado and Munk [1998] and Munk and Czado [1998] study the existence of relationships between distributions F and G by using a discrepancy measure between the distributions, $\Delta(F, G)$, built using the Wasserstein distance. The authors consider the assumption $\Delta(F, G) > \Delta_0$ versus $\Delta(F, G) \leq \Delta_0$ for Δ_0 a chosen threshold. Thus when the test is rejected, this implies that there is a statistical evidence that the two distributions are similar with respect to the chosen criterion. In this direction, we define a notion of variation of distributions using the Wasserstein distance, $\mathcal{W}_r$, in the set of probability measures with finite r-th moments, $\mathcal{F}_r(\mathbb{R}^d)$, $r \geq 1$, which generalizes the notion of variance for random distributions over $\mathbb{R}^d$. This

quantity can be defined as

$$V_r(\mu_1, \ldots, \mu_J) = \inf_{\eta \in \mathcal{F}_r(\mathbb{R}^d)} \left(\frac{1}{J} \sum_{j=1}^{J} \mathcal{W}_r^r(\mu_j, \eta) \right)^{1/r},$$

which measures the spread of the distributions. Then, to measure closeness to a deformation model we take a look at the minimal variation among warped distributions, a quantity that we could consider as a minimal alignment cost. Under some mild conditions a deformation model holds if and only if this minimal alignment cost is null and we can base our assessment of a deformation model on this quantity. As in Czado and Munk [1998] and Munk and Czado [1998] we provide results (CLT's and bootstrap versions) that enable to reject that the minimal alignment cost exceeds some threshold (hence, to conclude that it is below that threshold). Our results are given in a setup of general, nonparametric classes of warping functions. If, still, one is interested in the more classical goodness-of-fit problem for the deformation model we also provide results in a somewhat more restrictive setup.

The paper is organized as follows. The main facts about Wasserstein variation are presented in Section 2, together with the key idea that fit to a deformation model can be recast in terms of the minimal Wasserstein variation among warped versions of the distributions. Later, in Section 3 we prove some Lipsichtz bounds for the law of empirical Wasserstein variations as well as of minimal alignment costs on $\mathbb{R}^d$. The implications of these results include that quantiles of the minimal warped variation criterion can be consistently estimated by some suitable bootstrap quantiles, which can be approximated by simulation, yielding some consistent tests of fit to deformation models, provided that the empirical criterion has some regular limiting distribution. This issue, namely, Central Limit Theorems for empirical minimal Wasserstein variation is further explored for univariate distributions in Sections 4, covering non parametric deformation models, and 5, with a sharper analysis for the case of semiparametric deformation models. These sections propose consistent tests for deformation models in the corresponding setups. Section 6 provides some simulations to assess the quality of the bootstrap procedure. Finally, proofs are postponed to Section 7.

7.2 Wasserstein variation and deformation models for distributions

Much recent work has been conducted to measure the spread or the inner structure of a collection of distributions. In this paper we define a notion of variability which relies on the notion of Fréchet mean for the space of probability endowed with the Wasserstein metrics, of which we will recall the definition hereafter. First, for $d \geq 1$, consider the set $\mathcal{F}_r\left(\mathbb{R}^d\right)$ of probabilities with finite r-th moment. For μ and ν in $\mathcal{F}_r\left(\mathbb{R}^d\right)$, we denote by $\Pi(\mu, \nu)$ the set of all probability measures π over the product set $\mathbb{R}^d \times \mathbb{R}^d$ with first (resp. second) marginal μ (resp. ν). The L_r transportation cost between these two measures is defined as

$$\mathcal{W}_r(\mu, \nu)^r = \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^r \, d\pi(x, y).$$

This transportation cost allows to endow the set $\mathcal{F}_r\left(\mathbb{R}^d\right)$ with the metric $\mathcal{W}_r(\mu, \nu)$. More details on Wasserstein distances and their links with optimal transport problems can be found in Rachev [1984] or Villani [2009] for instance.

Within this framework, we can define a global measure of separation of a collection of measures μ_j, $j = 1, \ldots, n$, as follows. Given probabilities $\mu_1, \ldots, \mu_J \in \mathcal{F}_r(\mathbb{R}^d)$ let

$$V_r(\mu_1, \ldots, \mu_J) = \inf_{\eta \in \mathcal{F}_r(\mathbb{R}^d)} \Big(\frac{1}{J} \sum_{j=1}^{J} \mathcal{W}_r^r(\mu_j, \eta) \Big)^{1/r}$$

be the Wasserstein r-variation of $\mu_1, \ldots, \mu_J$ or the variance of the μ_j's.

The special case $r = 2$ has been studied in the literature. Existence of a minimizer of the map $\eta \mapsto \frac{1}{J} \sum_{j=1}^{J} \mathcal{W}_2^2(\mu_j, \eta)$ is proved in Agueh and Carlier [2011], as well as uniqueness under some smoothness assumptions. Such a minimizer, μ_B, is called a barycenter or Fréchet mean of $\mu_1, \ldots, \mu_J$. Hence, $V_2(\mu_1, \ldots, \mu_J) = (\frac{1}{J} \sum_{j=1}^{J} \mathcal{W}_2^2(\mu_j, \mu_B))^{1/2}$. Empirical versions of the barycenter are analyzed in Boissard et al. [2015] or Le Gouic and Loubes [2017]. Similar ideas have also been developed in Cuturi and Doucet [2014] or Bigot and Klein [2018].

This quantity, which is an extension of the variance for probability distributions is a good candidate to evaluate the concentration of a collection of measures around its Fréchet mean. In particular, it can be used to measure fit to a distribution deformation model. More precisely, assume as in the Introduction that we observe J independent i.i.d. samples with sample j, $j = 1, \ldots, J$ consisting of i.i.d. observations $X_{i,j}$, $i = 1, \ldots, n$ with common distribution μ_j. We change for later convenience the notation in the Introduction. We assume that $\mathcal{G}_j$ is a family (parametric or nonparametric) of invertible warping functions and denote $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_J$. The deformation model assumes then that

$$\begin{gathered} \text{there exists } (\varphi_1^*, \ldots \varphi_J^*) \in \mathcal{G} \text{ and i.i.d. } (\varepsilon_{i,j})_{\substack{1 \le i \le n \\ 1 \le j \le J}} \text{ such that} \\ X_{i,j} = \left(\varphi_j^*\right)^{-1} (\varepsilon_{i,j}) \quad \forall 1 \le j \le J \end{gathered} \tag{7.2.1}$$

Equivalently, the deformation model (7.2.1) means that there exist $(\varphi_1^*, \ldots \varphi_J^*) \in \mathcal{G}$ such that $\varphi_j^*(X_{i,j})$, $1 \le j \le J$, $1 \le i \le n$, are all i.i.d. or, if we write $\mu_j(\varphi_j)$ for the distribution of $\varphi_j(X_{i,j})$, that there exists $(\varphi_1^*, \ldots \varphi_J^*) \in \mathcal{G}$ such that

$$\mu_1(\varphi_1^*) = \cdots = \mu_J(\varphi_J^*). \tag{7.2.2}$$

We propose to use the Wasserstein variation to measure fit to model (7.2.1), through the minimal alignment cost

$$A_r(\mathcal{G}) := \inf_{(\varphi_1, \ldots, \varphi_J) \in \mathcal{G}} V_r^r \left(\mu_1(\varphi_1), \ldots, \mu_J(\varphi_J)\right). \tag{7.2.3}$$

Let us assume that $\mu_1(\varphi_1), \ldots, \mu_J(\varphi_J)$, $(\varphi_1, \ldots, \varphi_J) \in \mathcal{G}$ are in $\mathcal{F}_r(\mathcal{R}^d)$. If the deformation model (7.2.1) holds then $A_r(\mathcal{G}) = 0$. Under the additional mild assumption that the minimum in (7.2.3) is attained we have that the deformation model can be equivalently formulated as

$$A_r(\mathcal{G}) = 0 \tag{7.2.4}$$

and a goodness-of-fit test to the deformation model becomes, formally, a test of

$$H_0 : A_r(\mathcal{G}) = 0 \quad \text{vs.} \quad H_a : A_r(\mathcal{G}) > 0. \tag{7.2.5}$$

A testing procedure can be based on the empirical version of $A_r(\mathcal{G})$, namely,

$$A_{n,r}(\mathcal{G}) := \inf_{(\varphi_1, \ldots, \varphi_J) \in \mathcal{G}} V_r^r \left(\mu_{n,1}(\varphi_1), \ldots, \mu_{n,J}(\varphi_J)\right), \tag{7.2.6}$$

where $\mu_{n,j}(\varphi_j)$ denotes the empirical measure on $\varphi_j(X_{1,j}), \ldots, \varphi_j(X_{n,j})$. We would reject the deformation model (7.2.1) for large values of $A_{n,r}(\mathcal{G})$.

As noted in Czado and Munk [1998] or Munk and Czado [1998] the testing problem (7.2.5) can be considered as a mere sanity check for the deformation model, since lack of rejection of the null does not provide statistical evidence that the deformation model holds. Consequently, as in the cited references, we will also consider the alternative testing problem

$$H_0 : A_r(\mathcal{G}) \geq \Delta_0 \quad \text{vs.} \quad H_a : A_r(\mathcal{G}) < \Delta_0, \tag{7.2.7}$$

where $\Delta_0 > 0$ is a fixed threshold. With this formulation the test decision of rejecting the null hypothesis implies that there is statistical evidence that the deformation model is approximately true. In this case rejection would correspond to small observed values of $A_{n,r}(\mathcal{G})$. In later sections we provide theoretical results that allow the computation of approximate critical values and p-values for the testing problems (7.2.5) and (7.2.7) under suitable assumptions.

7.3 Bootstraping Wasserstein's variations

We present now some general results on Wasserstein distances that will be applied to estimate the asymptotic distribution of the minimal alignment cost statistic, $A_{n,r}(\mathcal{G})$, defined in (7.2.6). In this section, we write $\mathcal{L}(Z)$ for the law of any random variable Z. We note the abuse of notation in the following, in which W_r is used both for Wasserstein distance on $\mathbb{R}$ and on $\mathbb{R}^d$, but this should not cause much confusion.

Our first result shows that the laws of empirical transportation costs are continuous (and even Lipschitz) functions of the underlying distributions.

Theorem 7.3.1 *Set ν, ν', η probability measures in $\mathcal{F}_r\left(\mathbb{R}^d\right)$, $Y_1, \ldots, Y_n$ i.i.d. random vectors with common law ν, $Y'_1, \ldots, Y'_n$, i.i.d. with law ν' and write ν_n, ν'_n for the corresponding empirical measures. Then*

$$\mathcal{W}_r(\mathcal{L}(\mathcal{W}_r(\nu_n, \eta)), \mathcal{L}(\mathcal{W}_r(\nu'_n, \eta))) \leq W_r(\nu, \nu').$$

The deformation assessment criterion introduced in section 2 is basd on the Wasserstein r-variation of distributions, V_r. It is convenient to note that $V_r^r(\nu_1, \ldots, \nu_J)$ can also be expressed as

$$V_r^r(\nu_1, \ldots, \nu_J) = \inf_{\pi \in \Pi(\nu_1, \ldots, \nu_J)} \int T(y_1, \ldots, y_J) d\pi(y_1, \ldots, y_J), \tag{7.3.1}$$

where $\Pi(\nu_1, \ldots, \nu_J)$ denotes the set of probability measures on $\mathbb{R}^d$ with marginals $\nu_1, \ldots, \nu_J$ and $T(y_1, \ldots, y_J) = \min_{z \in \mathbb{R}^d} \frac{1}{J} \sum_{j=1}^J \|y_j - z\|^r$.

Here we are interested in empirical Wasserstein r-variations, namely, the r-variations computed from the empirical measures $\nu_{n_j,j}$ coming from independent samples $Y_{1,j}, \ldots, Y_{n_j,j}$ of i.i.d. random variables with distribution ν_j. Note that in this case problem (7.3.1) is a linear optimization problem for which a minimizer always exists.

As before, we consider the continuity of the law of empirical Wasserstein r-variations with respect to the underlying probabilities. This is covered in the next result.

Theorem 7.3.2 *With the above notation*

$$\mathcal{W}_r^r(\mathcal{L}(V_r(\nu_{n_1,1}, \ldots, \nu_{n_J,J})), \mathcal{L}(V_r(\nu'_{n_1,1}, \ldots, \nu'_{n_J,J}))) \leq \frac{1}{J} \sum_{j=1}^J \mathcal{W}_r^r(\nu_j, \nu'_j).$$

A useful consequence of the above results is that empirical Wasserstein distances or r-variations can be bootstrapped under rather general conditions. To be more precise, we take in Theorem 7.3.1 $\nu' = \nu_n$, the empirical measure on $Y_1, \ldots, Y_n$ and consider a bootstrap sample $Y_1^*, \ldots, Y_{m_n}^*$ of i.i.d. (conditionally given $Y_1, \ldots, Y_n$) observations with common law ν_n. We will assume that the resampling size m_n satisfies $m_n \to \infty$, $m_n = o(n)$ and write $\nu_{m_n}^*$ for the empirical measure on $Y_1^*, \ldots, Y_{m_n}^*$ and $\mathcal{L}^*(Z)$ for the conditional law of Z given $Y_1, \ldots, Y_n$. Theorem 7.3.1 now reads

$$\mathcal{W}_r(\mathcal{L}^*(\mathcal{W}_r(\nu_{m_n}^*, \nu)), \mathcal{L}(\mathcal{W}_r(\nu_{m_n}, \nu))) \leq \mathcal{W}_r(\nu_n, \nu).$$

Hence, if $\mathcal{W}_r(\nu_n, \nu) = O_{\mathbb{P}}(1/r_n)$ for some sequence $r_n > 0$ such that $r_{m_n}/r_n \to 0$ as $n \to \infty$, then, using that $\mathcal{W}_r(\mathcal{L}(aX), \mathcal{L}(aY)) = a\mathcal{W}_r(\mathcal{L}(X), \mathcal{L}(Y))$ for $a > 0$, we see that

$$\mathcal{W}_r(\mathcal{L}^*(r_{m_n}\mathcal{W}_r(\nu_{m_n}^*, \nu)), \mathcal{L}(r_{m_n}\mathcal{W}_r(\nu_{m_n}, \nu))) \leq \frac{r_{m_n}}{r_n} r_n \mathcal{W}_r(\nu_n, \nu) \to 0 \tag{7.3.2}$$

in probability.

Asume that, in addition, $r_n\mathcal{W}_r(\nu_n, \nu) \rightharpoonup \gamma(\nu)$ for a smooth distribution $\gamma(\nu)$. Then (see, e.g., Lemma 1 in Janssen and Pauls [2003]) if $\hat{c}_n(\alpha)$ denotes the α quantile of the conditional distribution $\mathcal{L}^*(r_{m_n}\mathcal{W}_r(\nu_{m_n}^*, \nu))$

$$\mathbb{P}\left(r_n\mathcal{W}_r(\nu_n, \nu) \leq \hat{c}_n(\alpha)\right) \to \alpha \quad \text{as } n \to \infty. \tag{7.3.3}$$

We conclude in this case that the quantiles of $r_n\mathcal{W}_r(\nu_n, \nu)$ can be consistently estimated by the bootstrap quantiles, $\hat{c}_n(\alpha)$, which, in turn, can be approximated through Monte-Carlo simulation.

As an example, if $d = 1$ and $r = 2$, under integrability and smoothness assumptions on ν we have $\sqrt{n}\mathcal{W}_2(\nu_n, \nu) \rightharpoonup \left(\int_0^1 \frac{B^2(t)}{f^2(F^{-1}(t))}dt\right)^{1/2}$, where f and F^{-1} are the density and the quantile function of ν, see del Barrio et al. [2005], and (7.3.3) holds.

For the deformation model (7.2.1), statistical inference is based on $A_{n,r}(\mathcal{G})$, introduced in (7.2.6). Now consider $A'_{n,r}(\mathcal{G})$, the corresponding version obtained from samples with underlying distributions μ'_j. Then, a version of Theorem 7.3.2 is valid for these minimal alignment costs, provided the deformation classes are uniformly Lipschitz, namely, under the assumption that

$$L_j := \sup_{x \neq y, \varphi_j \in \mathcal{G}_j} \frac{\|\varphi_j(x) - \varphi_j(x)\|}{\|x - y\|}, \quad j = 1, \ldots, J \tag{7.3.4}$$

are finite.

Theorem 7.3.3 *If $L = \max(L_1, \ldots, L_j) < \infty$, with L_j as in (7.3.4), then*

$$\mathcal{W}_r^r(\mathcal{L}((A_{n,r}(\mathcal{G}))^{1/r}), \mathcal{L}((A'_{n,r}(\mathcal{G}))^{1/r})) \leq L^r \frac{1}{J}\sum_{j=1}^{J} \mathcal{W}_r^r(\mu_j, \mu'_j).$$

Hence, the Wasserstein distance of the variance of two collections of distributions can be controlled using the distance between the distributions. The main consequence of this fact is that the minimal alignment cost can be also bootstrapped as soon as a distributional limit theorem exists for $A_{n,r}(\mathcal{G})$, as in the discussion above. In sections 4 and 5 below we present distributional results of this type in the one dimensional case. We note that, while general central limit theorems for the empirical transportation cost are not available in dimension $d > 1$, some recent progress has been made in this line, see, e.g., Rippl et al. [2016] for Gaussian distributions and Sommerfeld and Munk [2018], which gives such type of results for distributions on $\mathbb{R}^d$ with finite support. Further advances in this line would enable to extend the results in the following section to higher dimension.

7.4 Assessing fit to non-parametric deformation models

We focus in this and the next sections on the case $d = 1$ and $r = 2$ and will simply write $A(\mathcal{G})$ and $A_n(\mathcal{G})$ (instead of $A_2(\mathcal{G})$ and $A_{2,n}(\mathcal{G})$) for the minimal alignment cost and its empirical version, defined in (7.2.3) and (7.2.6). Otherwise we keep the notation in section 2, with $X_{1,j}, \ldots, X_{n,j}$ i.i.d. r.v.s with law μ_j being one of the J independent samples. Now $\mathcal{G}_j$ is a class of invertible warping functions from $\mathbb{R}$ to $\mathbb{R}$ which we assume to be increasing. We note that in this case the barycenter of a set of probabilities $\mu_1, \ldots, \mu_J$ with distribution functions $F_1, \ldots, F_J$ is the probability having quantile function $F_B^{-1} := \frac{1}{J}\sum_{j=1}^{J} F_j^{-1}$, see, e.g., Agueh and Carlier [2011]. We observe further that $\mu_j(\varphi_j)$ is determined by the quantile function $\varphi_j \circ F_j^{-1}$. We will write

$$F_B^{-1}(\varphi) = \frac{1}{J}\sum_{j=1}^{J} \varphi_j \circ F_j^{-1} \tag{7.4.1}$$

for the quantile function of the barycenter of $\mu_1(\varphi_1), \ldots, \mu_J(\varphi_J)$, while $\rightharpoonup$ will denote convergence in distribution.

In order to prove a CLT for $A_n(\mathcal{G})$ we need to make assumptions on the integrability and regularity of the distributions μ_j as well as on the smoothness of the warping functions. We consider first the assumptions on the distributions. For each μ_j, $j = 1, \ldots, J$, we denote its distribution function by F_j. We will assume that μ_j is supported on an (possibly unbounded) interval in the interior of which F_j is C^2 and $F_j' = f_j > 0$ and satisfies

$$\sup_x \frac{F_j(x)(1-F_j(x))f_j'(x)}{f_j(x)^2} < \infty, \tag{7.4.2}$$

and, further, that for some $q > 1$

$$\int_0^1 \frac{(t(1-t))^{\frac{q}{2}}}{\left(f_j\left(F_j^{-1}(t)\right)\right)^q} dt < \infty \tag{7.4.3}$$

and for some $r > 4$

$$\mathbb{E}\left[|X_j|^r\right] < \infty. \tag{7.4.4}$$

Assumption (7.4.2) is a classical regularity requirement for the use of strong approximations for the quantile process, as in Csörgö and Horváth [1993] or del Barrio et al. [2005]. Our proof relies on the use of these techniques. Then (7.4.3) and (7.4.4) are mild integrability conditions. If F_j has regularly varying tails of order $-r$ (as, for instance, Pareto tails) then both conditions hold (and also (7.4.2)) as long as $r > 4$ and $1 < q < 2r/(r+2)$. Of course the conditions are fulfilled by distributions with lighter tails such as exponential or Gaussian laws (for any $q \in (1, 2)$).

Turning to the assumptions on the classes of warping functions, we recall that a uniform Lipsichtz condition was needed for the approximation bound in Theorem 7.3.3. For the CLT in this section we need some refinement of that condition, the extent of which will depend on the integrability exponent q in (7.4.3), as follows. We set $p_0 = \max\left(\frac{q}{q-1}, 2\right)$ and define on $\mathcal{H}_j = C^1(\mathbb{R}) \cap L^{p_0}(X_j)$ the norm $\|h_j\|_{\mathcal{H}_j} = \sup|h_j'(x)| + \mathbb{E}\left[|h_j(X_j)|^{p_0}\right]^{\frac{1}{p_0}}$, and on the product space $\mathcal{H}_1 \times \cdots \times \mathcal{H}_J$, $\|h\|_{\mathcal{H}} = \sum_{j=1}^{J} \|h_j\|_{\mathcal{H}_j}$ and assume that

$$\mathcal{G}_j \subset \mathcal{H}_j \text{ is compact for } \|\cdot\|_{\mathcal{H}_j} \text{ and } \sup_{h\in\mathcal{G}_j} \left|h'(x_n^h) - h'(x)\right| \underset{\sup_{h\in\mathcal{G}_j}|x_n^h - x|\to 0}{\longrightarrow} 0, \tag{7.4.5}$$

and, finally, that for some $r > \max(4, p_0)$,

$$\mathbb{E} \sup_{h \in \mathcal{G}_j} |h(X_j)|^r < \infty. \tag{7.4.6}$$

We note that (7.4.6) is a slight strengthening of the uniform moment bound already contained in (7.4.5) (we could take $p_0 > \max(\frac{q}{q-1}, 4)$ in (7.4.5) and (7.4.6) would follow). Our next result gives a CLT for $A_n(\mathcal{G})$ under the assumptions on the distributions and deformation classes described above. The limit can be simply described in terms of a centered Gaussian process indexed by the set of minimizers of the variation functional, namely,

$$U(\varphi) = V_2^2(\mu_1(\varphi_1), \ldots, \mu_J(\varphi_J)).$$

An elementary computation shows that $(U^{1/2}(\varphi) - U^{1/2}(\tilde{\varphi}))^2 \leq \frac{1}{J} \sum_{j=1}^J \mathbb{E}(\varphi_j(X_j) - \tilde{\varphi}_j(X_j))^2$, from which we conclude continuity of U with respect to $\|\cdot\|_{\mathcal{H}}$. In particular, the set

$$\Gamma = \left\{ \varphi \in \mathcal{G} : U(\varphi) = \inf_{\phi \in \mathcal{G}} U(\phi) \right\} \tag{7.4.7}$$

is a nonempty compact subset of $\mathcal{G}$.

Theorem 7.4.1 *Assume that $(B_j)_{1 \leq j \leq J}$ are independent Brownian bridges. Set*

$$c_j(\varphi) = 2 \int_0^1 \varphi_j' \circ F_j^{-1} (\varphi_j \circ F_j^{-1} - F_B^{-1}(\varphi)) \frac{B_j}{f_j \circ F_j^{-1}}$$

and $C(\varphi) = \frac{1}{J} \sum_{j=1}^J c_j(\varphi)$, $\varphi \in \mathcal{G}$. Then, under assumptions (7.4.2) to (7.4.6), C is a centered Gaussian process on $\mathcal{G}$ with trajectories a.s. continuous with respect to $\|\cdot\|_{\mathcal{H}}$. Furthermore,

$$\sqrt{n}(A_n(\mathcal{G}) - A(\mathcal{G})) \rightharpoonup \min_{\varphi \in \Gamma} C(\varphi).$$

A proof of Theorem 7.4.1 is given in the Appendix below. The random variables $\int_0^1 \varphi_j' \circ F_j^{-1} \frac{B_j}{f_j \circ F_j^{-1}} (\varphi_j \circ F_j^{-1} - F_B^{-1}(\varphi))$ are centered Gaussian, with variance

$$\int_{[0,1]^2} (\min(s,t) - st) \frac{\varphi_j'(F_j^{-1}(t))}{f_j(F_j^{-1}(t))} (\varphi_j(F_j^{-1}(t)) - F_B^{-1}(\varphi)(t)) \\ \times \quad \frac{\varphi_j'(F_j^{-1}(s))}{f_j(F_j^{-1}(s))} (\varphi_j(F_j^{-1}(s)) - F_B^{-1}(\varphi)(s)) ds dt.$$

In particular, if U has a unique minimizer the limiting distribution in Theorem 7.4.1 is normal. However, our result works in more generality, even without uniqueness assumptions.

We remark also that although we have focused for simplicity on the case of samples of equal size, the case of different sample sizes, n_j, $j = 1, \ldots, J$, can also be handled with straightforward changes. More precisely, let us write $A_{n_1, \ldots, n_J}(\mathcal{G})$ for the minimal alignment cost computed from the empirical distribution of the samples and assume that $n_j \to +\infty$ and

$$\frac{n_j}{n_1 + \cdots + n_J} \to (\gamma_j)^2 > 0,$$

then with straightforward changes in our proof we can see that

$$\sqrt{\frac{n_1 \ldots n_J}{(n_1 + \cdots + n_J)^{J-1}}} (A_{n_1, \ldots, n_J}(\mathcal{G}) - A(\mathcal{G})) \rightharpoonup \min_{\varphi \in \Gamma} \tilde{C}(\varphi), \tag{7.4.8}$$

where $\tilde{C}(\varphi) = \frac{1}{J}\sum_{j=1}^{J} \tilde{c}_j(\varphi)$ and $\tilde{c}_j(\varphi) = \big(\Pi_{p\neq j}\gamma_p\big)c_j(\varphi)$.

If we try, as argued in section 2, to base our assessment of fit to the deformation model (7.2.1) on $A_n(\mathcal{G})$, we should note that the limiting distribution in Theorem 7.4.1 depends on the unknown distributions μ_j and cannot be used for the computation of approximate critical values or p-values without further adjustments. We show now how this can be done in the case of the testing problem (7.2.7), namely, the test of

$$H_0 : A_r(\mathcal{G}) \geq \Delta_0 \quad \text{vs.} \quad H_a : A_r(\mathcal{G}) < \Delta_0,$$

for some fixed threshold $\Delta_0 > 0$, through the use of a bootstrap procedure.

Let us consider bootstrap samples $X^*_{1,j}, \ldots, X^*_{m_n,j}$ of i.i.d. observations sampled from $\mu_{n,j}$, the empirical distribution on $X_{1,j}, \ldots, X_{n,j}$. We write $\mu^*_{m_n,j}$ for the empirical measure on $X^*_{1,j}, \ldots, X^*_{m_n,j}$ and introduce

$$A^*_{m_n}(\mathcal{G}) = \inf_{\varphi\in\mathcal{G}} V_2^2(\mu^*_{m_n,1}(\varphi_1), \ldots, \mu^*_{m_n,J}(\varphi_J)).$$

Now, we base our testing procedure on the conditional α-quantiles (given the $X_{i,j}$'s) of $\sqrt{m_n}(A^*_{m_n}(\mathcal{G}) - \Delta_0)$, which we denote $\hat{c}_n(\alpha;\Delta_0)$. Our next result, which follows from Theorems 7.3.3 and 7.4.1, shows that the test that rejects H_0 when

$$\sqrt{n}(A_n(\mathcal{G}) - \Delta_0) < \hat{c}_n(\alpha;\Delta_0)$$

is a consistent test of approximate level α for (7.2.7). We note that the bootstrap quantiles $\hat{c}_n(\alpha;\Delta_0)$ can be computed using Monte-Carlo simulation.

Corollary 7.4.2 *If $m_n \to \infty$, and $m_n = O(\sqrt{n})$, then under assumptions (7.4.2) to (7.4.6)*

$$\mathbb{P}\big(\sqrt{n}(A_n(\mathcal{G}) - \Delta_0) < \hat{c}_n(\alpha;\Delta_0)\big) \to \begin{cases} 0 & \text{if } A(\mathcal{G}) > \Delta_0 \\ \alpha & \text{if } A(\mathcal{G}) = \Delta_0 \\ 1 & \text{if } A(\mathcal{G}) < \Delta_0 \end{cases} \tag{7.4.9}$$

Rejection in the testing problem (7.2.7) would result, as noted in section 2, in statistical evidence supporting that the deformation model holds approximately (hence, that related registration methods can be safely applied). If, nevertheless, we were interested in gathering statistical evidence against the deformation model then we should consider the classical goodness-of-fit problem (7.2.5). Some technical difficulties arise then. Note that if the deformation model holds, that is, if $A(\mathcal{G}) = 0$, then we have $\varphi_j \circ F_j^{-1} = F_B^{-1}(\varphi)$ for each $\varphi \in \Gamma$, which implies that the result of Theorem 7.4.1 becomes

$$\sqrt{n}A_n(\mathcal{G}) \rightharpoonup 0.$$

Hence, a nondegenerate limit law for $A_n(\mathcal{G})$ in this case requires a more refined analysis, that we handle in the next section.

7.5 Goodness-of-fit in semiparametric deformation models

In many cases, deformation functions can be made more specific in the sense that they follow a known shape depending on parameters that may differ for sample to sample. In our approach to the classical goodness-of-fit problem (7.2.5) we consider a parametric model in which $\varphi_j = \varphi_{\theta_j}$ for some finite dimensional parameter θ_j that describes the warping effect within a fixed shape.

Now, that the deformation model holds means that there exist $\theta^* = (\theta_1^*, \ldots, \theta_J^*)$ such that for $1 \leq i \leq n, 1 \leq j \leq J$,

$$X_{i,j} = \varphi_{\theta_j^*}^{-1}(\varepsilon_{i,j}).$$

Hence, from now on, we will consider the following family of deformations, indexed by a parameter $\lambda \in \Lambda \subset \mathbb{R}^p$:

$$\begin{array}{rcl} \varphi : \Lambda \times \mathbb{R} & \to & \mathbb{R} \\ (\lambda, x) & \mapsto & \varphi_\lambda(x) \end{array}$$

The classes $\mathcal{G}_j$ become now $\{\varphi_{\theta_j} : \theta_j \in \Lambda\}$. We denote $\Theta = \Lambda^J$ and write $A_n(\Theta)$ and $A(\Theta)$ instead of $A_n(\mathcal{G})$ and $A(\mathcal{G})$. We also use the simplified notation $\mu_j(\theta_j)$ instead of $\mu_j(\varphi_{\theta_j})$, $F_B(\theta)$ for $F_B(\varphi_{\theta_1}, \ldots, \varphi_{\theta_J})$ and similarly for the empirical versions. Our main goal is to prove a weak limit theorem for $A_n(\Theta)$ under the null in (7.2.5). Therefore, throughout this section *we assume that model (7.2.1) holds*. This means, in particular, that the quantile functions of the samples satisfy $F_j^{-1} = \varphi_{\theta_j^*}^{-1} \circ G^{-1}$, with G the d.f. of the $\varepsilon_{i,j}$'s. As before, we assume that the warping functions are invertible and increasing, which now means that, for each $\lambda \in \Lambda$, φ_λ is an invertible, increasing function. It is convenient at this point to introduce the notation

$$\psi_j(\lambda, x) = \varphi_\lambda(\varphi_{\theta_j^*}^{-1}(x)), \quad j = 1, \ldots, J \tag{7.5.1}$$

and ε for a random variable with the same distribution as the $\varepsilon_{i,j}$. Note that $\psi_j(\theta_j^*, x) = x$.

Now, under smoothness assumptions on the functions ψ_j that we present in detail below, if the parameter space is compact then the function

$$U_n(\theta_1, \ldots, \theta_J) = V_2^2(\mu_{n,1}(\theta_1), \ldots, \mu_{n,J}(\theta_J))$$

admits a minimizer, that we will denote by $\hat{\theta}_n$, that is

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} U_n(\theta). \tag{7.5.2}$$

Of course, since we are assuming that the deformation model holds, we know that θ^* is a minimizer of

$$U(\theta_1, \ldots, \theta_J) = V_2^2(\mu_1(\theta_1), \ldots, \mu_J(\theta_J)).$$

For a closer analysis of the asymptotic behavior of $A_n(\Theta)$ under the deformation model we need to make the following identifiability assumption

$$\theta^* \text{ belongs to the interior of } \Lambda \text{ and is the unique minimizer of } U. \tag{7.5.3}$$

Note that, equivalently, this means that θ^* is the unique zero of U.

As in the case of nonparametric deformation models, we need to impose some conditions on the class of warping functions and on the distribution of the errors, the $\varepsilon_{i,j}$. For the former, we write D or D_u for derivative operators with respect to parameters (hence, for instance, $D\psi_j(\lambda, x) = (D_1\psi_j(\lambda, x), \ldots, D_p\psi_j(\lambda, x))^T$ is the vector consisting of partial derivatives of ψ_j with respect to its first p arguments evaluated at (λ, x); $D^2\psi_j(\lambda, x) = (D_{u,v}\psi_j(\lambda, x))_{u,v}$ is the hessian matrix for fixed x and so on). $\psi_j'(\lambda, x)$ and similar notation will stand for derivatives with respect to x. Then we will assume that for each $j = 1, \ldots, J$, $u, v = 1, \ldots, p$, and some $r > 4$

$$\psi_j(\cdot, \cdot) \text{ is } C^2, \tag{7.5.4}$$

$$\mathbb{E}\big[\sup_{\lambda \in \Lambda} |\psi_j(\lambda, \varepsilon)|^r\big] < \infty, \quad \mathbb{E}\big[\sup_{\lambda \in \Lambda} |D_u\psi_j(\lambda, \varepsilon)|^r\big] < \infty, \quad \mathbb{E}\big[\sup_{\lambda \in \Lambda} |D_{u,v}\psi_j(\lambda, \varepsilon)|^r\big] < \infty, \tag{7.5.5}$$

and

$$\psi_j'(\cdot,\cdot) \text{ is bounded on } \Lambda\times\mathbb{R} \quad \text{and} \quad \sup_{\lambda\in\Lambda}\left|\psi_j'(\lambda,x_n^\lambda)-\psi_j'(\lambda,x)\right| \xrightarrow{\sup_{\lambda\in\Lambda}|x_n^\lambda - x|\to 0} 0. \tag{7.5.6}$$

Turning to the distribution of the errors, we will assume that G is C^2 with $G'(x)=g(x)>0$ on some interval and

$$\sup_x \frac{G(x)\,(1-G(x))\,g'(x)}{g(x)^2} < \infty. \tag{7.5.7}$$

Additionally (but see the comments after Theorem 7.5.1 below) we make the assumption that

$$\int_0^1 \frac{t(1-t)}{g^2\left(G^{-1}(t)\right)}dt < \infty. \tag{7.5.8}$$

Finally, before stating the asymptotic result for $A_n(\Theta)$, we introduce the $p\times p$ matrices

$$\Sigma_{i,i} = \frac{2(J-1)}{J^2}\int_0^1 D_i\psi_i(\theta_i^*, G^{-1}(t))\psi_i(\theta_i^*, G^{-1}(t))^T dt,$$

$$\Sigma_{i,j} = -\frac{2}{J^2}\int_0^1 D_i\psi_i(\theta_i^*, G^{-1}(t))\psi_i(\theta_j^*, G^{-1}(t))^T dt, \quad i\neq j$$

and the $(pJ)\times(pJ)$ matrix

$$\Sigma = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,J} \\ \vdots & & \vdots \\ \Sigma_{J,1} & \cdots & \Sigma_{J,J} \end{bmatrix}. \tag{7.5.9}$$

Σ is a symmetric, positive semidefinite matrix. To see this, consider $x_1,\ldots,x_J\in\mathbb{R}^p$ and $x^T=[x_1^T,\ldots,x_J^T]$ and note that

$$\begin{aligned} x'\Sigma x &= \frac{2}{J^2}\int_0^1 \Big(\sum_i (J-1)(x_i\cdot D_i\psi_i(\theta_i^*, G^{-1}(t)))^2 \\ &- 2\sum_{i<j}(x_i\cdot D_i\psi_i(\theta_i^*, G^{-1}(t)))(x_j\cdot D_j\psi_j(\theta_j^*, G^{-1}(t)))\Big)dt \\ &= \frac{2}{J^2}\int_0^1 \sum_{i<j}((x_i\cdot D_i\psi_i(\theta_i^*, G^{-1}(t))) - (x_j\cdot D_j\psi_j(\theta_j^*, G^{-1}(t))))^2 dt \geq 0. \end{aligned}$$

In fact, Σ is positive definite, hence invertible, apart from some degenerate cases, For instance, if $p=1$, Σ is invertible unless all the functions $D_i\psi_i(\theta_i^*, G^{-1}(t))$ are proportional.

We are ready now for the announced distributional limit theorem.

Theorem 7.5.1 *Assume that the deformation model holds. Under assumptions (7.5.3) to (7.5.7)*

$$\hat{\theta}_n \to \theta^*$$

in probability. If, in addition, Φ is invertible, then

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \rightharpoonup \Sigma^{-1}Y,$$

where $Y=(Y_1^T,\ldots,Y_J^T)^T$ with

$$Y_j = \frac{2}{J}\int_0^1 D\psi_j(\theta_j^*, G^{-1}(t))\frac{\tilde{B}_j(t)}{g(G^{-1}(t))}dt,$$

$\tilde{B}_j = B_j - \frac{1}{J}\sum_{k=1}^{J} B_k$ *and* $(B_j)_{1\leq j\leq J}$ *independent Brownian bridges. Furthermore, if (7.5.8) also holds, then*

$$nA_n(\Theta) \rightharpoonup \frac{1}{J}\sum_{j=1}^{J}\int_0^1 \Big(\frac{\tilde{B}_j}{g\circ G^{-1}}\Big)^2 - \frac{1}{2}Y^T\Sigma^{-1}Y.$$

We have to make a number of comments here. First, we note that, while, for simplicity, we have formulated Theorem 7.5.1 assuming that the deformation model holds, the CLT for $\hat{\theta}_n$ still holds (with some additional assumptions and changes in Φ) in the case when the model is false and θ^* is not the *true* parameter, but the one that gives the best (but imperfect) alignment. Since our focus here is the assessment of the deformation models we refrain from pursuing this issue.

Our second comment is about the indentifiability condition (7.5.3). At first sight it can seem to be too strong to be realistic. Actually, for some deformation models it could happen that $\varphi_\theta \circ \varphi_\eta = \varphi_{\theta * \eta}$ for some $\theta * \eta \in \Theta$. In this case, if $X_{i,j} = \varphi^{-1}_{\theta_j^*}(\varepsilon_{i,j})$ with $\varepsilon_{i,j}$ i.i.d., then, for any θ, $X_{i,j} = \varphi^{-1}_{\theta * \theta_j^*}(\tilde{\varepsilon}_{i,j})$ with $\tilde{\varepsilon}_{i,j} = \varphi_\theta(\varepsilon_{i,j})$ which are also i.i.d. and, consequently, $(\theta * \theta_1^*, \ldots, \theta * \theta_J^*)$ is also a zero of U. This applies, for instance, to location and scale models. A simple fix to this issue is to select one of the signals as the reference, say the J-th signal, and assume that θ_J^* is known (since it can be, in fact, chosen arbitrarily). The criterion function becomes then $\tilde{U}(\theta_1, \ldots, \theta_{J-1}) = U(\theta_1, \ldots, \theta_{J-1}, \theta_J^*)$. One could then make the (more realistic) assumption that $\tilde{\theta}^* = (\theta_1^*, \ldots, \theta_{J-1}^*)$ is the unique zero of $\tilde{U}$ and base the analysis on $\tilde{U}_n(\theta_1, \ldots, \theta_{J-1}) = U_n(\theta_1, \ldots, \theta_{J-1}, \theta_J^*)$ and $\hat{\tilde{\theta}}_n = \arg\min_{\tilde{\theta}} \tilde{U}_n(\tilde{\theta})$. The results in this section can be adapted almost verbatim to this setup. In particular, $\sqrt{n}(\hat{\tilde{\theta}}_n - \tilde{\theta}^*) \rightharpoonup \tilde{\Sigma}^{-1}\tilde{Y}$, with $\tilde{Y}^T = (Y_1^T, \ldots, Y_{J-1}^T)$ and $\tilde{\Sigma} = [\Sigma_{i,j}]_{1\leq i,j\leq J-1}$. Again, the invertibility of $\tilde{\Sigma}$ is almost granted. In fact, arguing as above, we see that and $\tilde{\Sigma}$ is positive definite if the functions $D\psi_i(\theta_i^*, G^{-1}(t))$, $i = 1, \ldots, J-1$, are not null.

Next, we discuss about the smoothness and integrability conditions on the errors. As before, (7.5.7) is a regularity condition that enables to use strong approximations for the quantile process. One might be surprised that the moment condition (7.4.4) does not show up here, but in fact it is contained in (7.5.5) (recall that $\psi_j(\theta_j^*, x) = x$). The integrability condition (7.5.8) is necessary and sufficient for ensuring $\int_0^1 \frac{B(t)^2}{g^2(G^{-1}(t))}dt < \infty$ (from which we see that the limiting random variable in the last claim in Theorem 7.5.1 is an a.s. finite random variable) and implies that

$$n\mathcal{W}_2^2(G_n, G) \rightharpoonup \int_0^1 \frac{B(t)^2}{g^2(G^{-1}(t))}dt,$$

with G_n the empirical d.f. on a sample of size n and d.f G. We refer to del Barrio et al. [2005] and Samworth and Johnson [2004] for details. Condition (7.4.4) is a strong assumption on the tails of G and does not include, for instance, normal distributions. On the other hand, under the less stringent condition

$$\int_0^1\int_0^1 \frac{(s\wedge t - st)^2}{g^2(G^{-1}(s))g^2(G^{-1}(t))}dsdt < \infty, \tag{7.5.10}$$

which is satisfied for normal laws, it can be shown that the limit as $\delta \to 0$

$$\int_\delta^{1-\delta} \frac{B(t)^2 - t(1-t)}{g^2(G^{-1}(t))}dt,$$

exists in probability and can be expressed as a weighted sum of independent, centered χ_1^2 random variables, see del Barrio et al. [2005] for details. Then, denoting that kind of limits as

$\int_0^1 \frac{B(t)^2 - t(1-t)}{g^2(G^{-1}(t))} dt$, under some additional tail conditions (still satisfied by normal laws; these are conditions (2.10) and (2.22) to (2.24) in the cited reference) we have

$$n\mathcal{W}_2^2(G_n, G) - c_n \rightharpoonup \int_0^1 \frac{B(t)^2 - t(1-t)}{g^2(G^{-1}(t))} dt,$$

with $c_n = \int_{1/n}^{1-1/n} \frac{EB(t)^2}{g^2(G^{-1}(t))} dt$. A simple look at the proof of Theorem 5.1 shows that under these conditions (instead of (7.5.8)) we can conclude that

$$nA_n(\Theta) - \tfrac{J-1}{J^2} c_n \rightharpoonup \tfrac{1}{J} \sum_{j=1}^J \int_0^1 \frac{\tilde{B}_j^2(t) - \frac{J-1}{J} t(1-t)}{g^2(G^{-1}(t))} dt - \tfrac{1}{2} Y^T \Sigma^{-1} Y. \tag{7.5.11}$$

Our last comment about the assumptions for Theorem 7.5.1 concerns the compactness assumption on the parameter space. This may lead in some examples to artificial constraints on the parameter space. On the other hand, under some conditions (see, e.g., Corollary 3.2.3 in Van der Vaart and Wellner) it is possible to prove that the global minimizer of the empirical criterion lies in a compact neighborhood of the true minimizer. In such cases the conclusion of Theorem 7.5.1 would extend for the unconstrained deformation model. As a toy example consider the case of deformations by changes in scale, with $J = 2$. As above we fix the parameters of, say, the first sample, and consider the family of deformations $\varphi_\sigma(x) = \sigma x$. We assume that the deformation model holds, with the first sample having d.f. G and the second $\frac{1}{\sigma^*} G^{-1}$ (hence, σ^* is the unique minimizer of $U(\sigma)$). We obtain that $U_n(\sigma) = \frac{1}{4} \int_0^1 (F_{n,1}^{-1} - \sigma F_{n,2}^{-1})^2$, from which we see that $\hat{\sigma}_n = \big(\int F_{n,1}^{-1} F_{n,2}^{-1} \big) / \big(\int (F_{n,2}^{-1})^2 \big) \to \sigma^*$ a.s. and thus the conclusion of Theorem 7.5.1 remains valid if we take $\Theta = (0, \infty)$. To avoid further technicalities we prefer to think of this as a different problem that should be handled in an ad hoc way for each particular example.

Turning back to our goal of assessment of the deformation model (7.2.1) based on the observed value of $A_n(\Theta)$, Theorem 7.5.1 gives some insight into the threshold levels for rejection of the null in the testing problem (7.2.5). However, the limiting distribution still depends on unknown objects and designing a tractable test requires to estimate the quantiles of this distribution. This is the goal of our next result.

We consider bootstrap samples $X_{1,j}^*, \ldots, X_{m_n,j}^*$ of i.i.d. observations sampled from μ_j^n, write $\mu_{m_n,j}^*$ for the empirical measure on $X_{1,j}^*, \ldots, X_{m_n,j}^*$ and $A_{m_n}^*(\Theta)$ for the minimal alignment cost computed from the bootstrap samples. We also write $\hat{c}_n(\alpha)$ for the conditional α quantile of $m_n A_{m_n}^*(\Theta)$ given the $X_{i,j}$.

Corollary 7.5.2 *Assume that the semiparametric deformation models holds. If $m_n \to \infty$, and $m_n/n \to 0$, then under assumptions (7.5.3) to (7.5.8) we have that*

$$\mathbb{P}\left(nA_n(\Theta) > \hat{c}_n(1-\alpha)\right) \to \alpha. \tag{7.5.12}$$

Corollary 7.5.2 show that the test that rejects $H_0 : A(\Theta) = 0$ (which, as disussed in section 2, is true if and only if the deformation model holds) when $nA_n(\Theta) > \hat{c}_n(1-\alpha)$ is asymptotically of level α. It is easy to check that the test is consistent against alternatives that satisfy regularity and integrability assumptions as in Theorem 7.5.1.

The key to Corollary 7.5.2 is that under the assumptions a bootstrap CLT holds for $m_n A_{m_n}^*(\Theta)$. As with Theorem 7.5.1, the integrability conditions on the errors can be relaxed and still have a bootstrap CLT. That would be the case if we replace (7.5.12) by (7.5.10) and the additional conditions mentioned above under which (7.5.11) holds. Then, the further assumption that the errors have a log-concave distribution and $m_n = O(n^\rho)$ for some $\rho \in (0, 1)$ would be enough to prove a bootstrap CLT, see the comments after the proof of Corollary 7.5.2 in the Appendix. In particular, a bootstrap CLT holds for Gaussian tails.

7.6 Simulations

We present in this section different simulations in order to study the goodness of fit test we propose in this paper. In this framework, we consider the scale-location family of deformations, i.e $\theta^* = (\mu^*, \sigma^*)$ and observations such that $X_{i,j} = \mu_j^* + \sigma_j^* \epsilon_{i,j}$, for different distributions of $\epsilon_{i,j}$.

7.6.1 Construction of an α-level test

First, we aim at studying the bootstrap procedure which enables to build the test. For this we choose a level $\alpha = 0.05$ and aim at estimating the quantile of the asymptotic distribution using a bootstrap method.

Let B be the number of bootstrap samples, we proceed as follows to design a bootstrapped goodness of fit test.

1. For all $b = 1, \dots, B$,
   1.1. For $j = 1, \dots, J$, create a bootstrap sample $X_{1,j}^{*b}, \dots, X_{m,j}^{*b}$, with fixed size $0 < m \leqslant n$, of the first observation sample $X_{1,j}, \dots, X_{n,j}$
   1.2. Compute $\left(u_m^{*b}\right)^2 = \inf_{\theta \in \Theta} U_m^{*b}(\theta)$.
2. Sort the values $\left(u_m^{*b}\right)^2, b = 1, \dots, B$,
$$\left(u_m^{*(1)}\right)^2 \leqslant \ldots \leqslant \left(u_m^{*(B)}\right)^2,$$
   then take $\hat{q}_m(1-\alpha) = u_m^{*(B(1-\alpha))}$, the $1-\alpha$ quantile of the bootstrap distribution of the statistic $\inf_{\theta \in \Theta} U_n(\theta)$.
3. The test rejects the null hypothesis if $n u_n^2 > m \left(u_m^{*(B(1-\alpha))}\right)^2$.

Once the test is built, we first ensure that the level of the test has been correctly achieved. For this we repeat the test for large K (here $K = 1000$) to estimate the probability of rejection of the test as
$$\hat{p}_n = \frac{1}{K} \sum_{k=1}^{K} 1_{\left(n u_{n,k}^2 > m \left(u_{m,k}^{*(B(1-\alpha))}\right)^2\right)}.$$
We present in Table 7.1 these results for different J and several choices for $m = m_n$ depending on the size of the initial sample.

As expected, the bootstrap method enables to build a test of level α provided the bootstrap sample is large enough. The required size of the sample increases with the number of different distributions J to be tested.

7.6.2 Power of the test procedure

Then we compute the power of previous test for several situations. In particular we must compute the probability of rejection of the null hypothesis under H_a. Hence for several number of distributions, we test the assumption that the model comes from a warping frame, when a different distribution called γ is observed. The simulations are conducted for the following choices of the number of sample and for the different distributions;

- $J = 2 : \mathcal{N}(0,1)$ and γ;
- $J = 3 : \mathcal{N}(0,1)$, $\mathcal{N}\left(5, 2^2\right)$ and γ;
- $J = 5 : \mathcal{N}(0,1)$, $\mathcal{N}\left(5, 2^2\right)$, $\mathcal{N}(3,1)$, $\mathcal{N}\left(1.5, 3^2\right)$ and γ;
- $J = 10 : \mathcal{N}(0,1)$, $\mathcal{N}\left(5, 2^2\right)$, $\mathcal{N}(3,1)$, $\mathcal{N}\left(1.5, 3^2\right)$, $\mathcal{N}\left(7, 4^2\right)$, $\mathcal{N}\left(2.5, 0.5^2\right)$, $\mathcal{N}\left(1, 1.5^2\right)$, $\mathcal{N}\left(4, 3^2\right)$, $\mathcal{N}\left(6, 5^2\right)$ and γ;

and also for different choices of γ.

- Exponential distribution with parameter 1;
- Double exponential with parameter 1 (a.k.a Laplace distribution);
- Student distribution $T(3)$ and $T(4)$ with 3 and 4 degrees of freedom.

All simulations are done for different sample sizes and different bootstrap samples, n and m_n. The results are presented in Tables 7.2, 7.3, 7.4 and 7.5, respectively.

We observe that the power of the test is very high in most of the cases. For the Exponential distribution, the power is close to 1. Indeed this distribution is very different from the Gaussian distribution since it is not symmetric, resulting easy to discard the null assumption. The three other distributions do share with the Gaussian the property of symmetry, and yet the power of the test is also close to one, increasing with the number of observations. Finally, for the Student's distribution, the higher the number of degrees of freedom, the more similar it becomes to a Gaussian distribution. This explains why it becomes more difficult for the test to reject the null hypothesis when using a Student with 4 degrees of freedom rather than with 3.

7.7 Appendix to Chapter 7

7.7.1 Proofs of section 7.3

Proof of Theorem 7.3.1. We set $T_n = \mathcal{W}_r(\nu_n, \eta)$ and $T'_n = \mathcal{W}_r(\nu'_n, \eta)$ and $\Pi_n(\eta)$ for the set of probabilities on $\{1, \ldots, n\} \times \mathbb{R}^d$ with first marginal equal to the discrete uniform distribution on $\{1, \ldots, n\}$ and second marginal equal to η and note that we have $T_n = \inf_{\pi \in \Pi_n(\eta)} a(\pi)$ if we denote

$$a(\pi) = \left(\int_{\{1,\ldots,n\} \times \mathbb{R}^d} \|Y_i - z\|^r d\pi(i,z) \right)^{1/r}.$$

We define similarly $a'(\pi)$ from the Y'_i sample to get $T'_n = \inf_{\pi \in \Pi_n(\eta)} a'(\pi)$. But then, using the inequality $|\|a\| - \|b\|| \leq \|a - b\|$,

$$|a(\pi) - a'(\pi)| \leq \left(\int_{\{1,\ldots,n\} \times \mathbb{R}^d} \|Y_i - Y'_i\|^r d\pi(i,z) \right)^{1/r} = \left(\frac{1}{n} \sum_{i=1}^{n} \|Y_i - Y'_i\|^r \right)^{1/r}$$

This implies that

$$|T_n - T'_n|^r \leq \frac{1}{n} \sum_{i=1}^{n} \|Y_i - Y'_i\|^r.$$

If we take now (Y, Y') to be an optimal coupling of ν and ν', so that $E\left[\|Y - Y'\|^r\right] = \mathcal{W}_r^r(\nu, \nu')$ and $(Y_1, Y_1'), \ldots, (Y_n, Y_n')$ to be i.i.d. copies of (Y, Y') we see that for the corresponding realizations of T_n and T_n' we have

$$\mathbb{E}\left[|T_n - T_n'|^r\right] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|Y_i - Y_i'\|^r\right] = \mathcal{W}_r(\nu, \nu')^r.$$

But this shows that $\mathcal{W}_r(\mathcal{L}(T_n), \mathcal{L}(T_n')) \leq \mathcal{W}_r(\nu, \nu')$, as claimed.

□

Proof of Theorem 7.3.2. We write $V_{r,n} = V_r(\nu_{n_1,1}, \ldots, \nu_{n_J,J})$ and $V_{r,n}' = V_r(\nu_{n_1,1}', \ldots, \nu_{n_J,J}')$. We note that

$$V_{r,n}^r = \inf_{\pi \in \Pi(U_1,\ldots,U_J)} \int T(i_1, \ldots, i_J) d\pi(i_1, \ldots, i_J),$$

where U_j is the discrete uniform distribution on $\{1, \ldots, n_j\}$ and

$$T(i_1, \ldots, i_J) = \min_{z \in \mathbb{R}^d} \frac{1}{J}\sum_{j=1}^{J}\|Y_{i_j,j} - z\|^r.$$

We write $T'(i_1, \ldots, i_J)$ for the equivalent function computed from the $Y_{i,j}'$'s. Hence we have

$$|T(i_1, \ldots, i_J)^{1/r} - T'(i_1, \ldots, i_J)^{1/r}|^r \leq \frac{1}{J}\sum_{j=1}^{J}\|Y_{i_j,j} - Y_{i_j,j}'\|^r,$$

which implies

$$\begin{aligned}
&\left|\left(\int T(i_1, \ldots, i_J) d\pi(i_1, \ldots, i_J)\right)^{1/r} - \left(\int T(i_1, \ldots, i_J) d\pi(i_1, \ldots, i_J)\right)^{1/r}\right|^r \\
&\quad\leq \int \frac{1}{J}\sum_{j=1}^{J}\|Y_{i_j,j} - Y_{i_j,j}'\|^r d\pi(i_1, \ldots, i_J) \\
&\quad= \frac{1}{J}\sum_{j=1}^{J}\int \|Y_{i_j,j} - Y_{i_j,j}'\|^r d\pi(i_1, \ldots, i_J) = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\|Y_{i,j} - Y_{i,j}'\|^r\right)
\end{aligned}$$

So,

$$|V_{r,n} - V_{r,n}'|^r \leq \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\|Y_{i,j} - Y_{i,j}'\|^r\right).$$

If we take (Y_j, Y_j') to be an optimal coupling of ν_j and ν_j' and $(Y_{1,j}, Y_{1,j}'), \ldots, (Y_{n_j,j}, Y_{n_j,j}')$ to be i.i.d. copies of (Y_j, Y_j'), for $j = 1, \ldots, J$, then we obtain

$$\mathbb{E}\left[|V_{r,n} - V_{r,n}'|^r\right] \leq \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\mathbb{E}\left[\|Y_{i,j} - Y_{i,j}'\|^r\right]\right) = \frac{1}{J}\sum_{j=1}^{J}\mathcal{W}_r^r(\nu_j, \nu_j').$$

The conclusion follows.

□

Proof of Theorem 7.3.3. We argue as in the proof of Theorem 7.3.2 and write

$$A_{n,r}(\mathcal{G}) = \inf_{\varphi \in \mathcal{G}}\left[\inf_{\pi \in \Pi(U_1,\ldots,U_J)} \int T(\varphi; i_1, \ldots, i_J) d\pi(i_1, \ldots, i_J)\right],$$

where $T(\varphi; i_1, \ldots, i_J) = \min_{y\in\mathbb{R}} \frac{1}{J}\sum_{j=1}^{J} \|Z_{i_j,j}(\varphi_j) - y\|^r$. We write $T'(\varphi; i_1, \ldots, i_J)$ for the same function computed on the $Z'_{i,j}(\varphi_j)$'s. Now, from the fact $\|Z_{i,j}(\varphi_j) - Z'_{i,j}(\varphi_j)\|^r \leq L^r \|X_{i,j} - X'_{i,j}\|^r$ we see that

$$|T(\varphi; i_1, \ldots, i_J)^{1/r} - T'(\varphi; i_1, \ldots, i_J)^{1/r}|^r \leq L^r \frac{1}{J}\sum_{j=1}^{J} \|X_{i_j,j} - X'_{i_j,j}\|^r$$

and, as a consequence, that

$$|V_r(\mu_{n,1}(\varphi_1), \ldots, \mu_{n,J}(\varphi_J)) - V_r(\mu'_{n,1}(\varphi_1), \ldots, \mu'_{n,J}(\varphi_J))|^r \leq \frac{L^r}{J}\sum_{j=1}^{J}\sum_{i_j=1}^{n_j} \frac{1}{n_j}\|X_{i_j,j} - X'_{i_j,j}\|^r$$

which implies

$$|(A_{n,r}(\mathcal{G}))^{1/r} - (A'_{n,r}(\mathcal{G}))^{1/r}|^r \leq \frac{L^r}{J}\sum_{j=1}^{J} \left(\frac{1}{n_j}\sum_{i=1}^{n_j} \|X_{i,j} - X'_{i,j}\|^r\right).$$

If, as in the proof of Theorem 7.3.2, we assume that $(X_{i,j}, X'_{i,j})$, $i = 1, \ldots, n_j$ are i.i.d. copies of an optimal coupling for μ_j and μ'_j, with different samples independent from each other we obtain that

$$\mathbb{E}\left[|(A_{n,r}(\mathcal{G}))^{1/r} - (A'_{n,r}(\mathcal{G}))^{1/r}|^r\right] \leq \frac{L^r}{J}\sum_{j=1}^{J} \mathcal{W}_r^r(\mu_j, \mu'_j).$$

□

7.7.2 Proofs of sections 7.4 and 7.5

We provide here proofs of the main results in sections 4 and 5. Our approach relies on the consideration the processes

$$C_n(\varphi) = \sqrt{n}(U_n(\varphi) - U(\varphi)) \quad \text{and} \quad C(\varphi) = \frac{1}{J}\sum_{j=1}^{J} c_j(\varphi), \quad \varphi \in \mathcal{G}, \tag{7.7.1}$$

where $U_n(\varphi) = V_2^2(\mu_{n,1}(\varphi_1), \ldots, \mu_{n,J}(\varphi_J))$, $U(\varphi) = V_2^2(\mu_1(\varphi_1), \ldots, \mu_J(\varphi_J))$,

$$c_j(\varphi) = 2\int_0^1 \varphi'_j \circ F_j^{-1}(\varphi_j \circ F_j^{-1} - F_B^{-1}(\varphi))\frac{B_j}{f_j \circ F_j^{-1}}$$

and $(B_j)_{1\leq j\leq J}$ are independent standard Brownian bridges on $(0,1)$. We prove below that the empirical deformation cost process C_n converges weakly to C as random elements in $L^\infty(\mathcal{G})$, the space of bounded, real valued functions on $\mathcal{G}$. Theorem 7.4.1 will follow as a corollary of this result.

We will make frequent use in this section of the following technical Lemma, which follows easily from the triangle and Holder's inequalities. We omit the proof.

Lemma 7.7.1 *Under Assumption (7.4.6)*

i) $\sup_{\varphi_j\in\mathcal{G}_j} \sqrt{n}\int_0^{\frac{1}{n}} (\varphi_j \circ F_j^{-1})^2 \to 0$, $\sup_{\varphi_j\in\mathcal{G}_j} \sqrt{n}\int_{1-\frac{1}{n}}^1 (\varphi_j \circ F_j^{-1})^2 \to 0$.

ii) $\sup_{\varphi_j\in\mathcal{G}_j} \sqrt{n}\int_0^{\frac{1}{n}} (\varphi_j \circ F_{n,j}^{-1})^2 \to 0$, $\sup_{\varphi_j\in\mathcal{G}_j} \sqrt{n}\int_{1-\frac{1}{n}}^1 (\varphi_j \circ F_{n,j}^{-1})^2 \to 0$ *in probability.*

iii) If moreover (7.4.3) holds then for all $1 \leq j,k \leq J$

$$\int_0^1 \frac{\sqrt{t(1-t)}}{f_k(F_k^{-1}(t))} \sup_{\varphi_j \in \mathcal{G}_j} \Big|\varphi_j(F_j^{-1}(t))\Big| dt < \infty \tag{7.7.2}$$

Theorem 7.7.2 *Under assumptions (7.4.2) to(7.4.6)* C_n *and* C *have a.s. trajectories in* $L^\infty(\mathcal{G})$. *Furthermore,* C *is a tight Gaussian random elemnt and* C_n *converges weakly to* C *in* $L^\infty(\mathcal{G})$.

Proof. We start noting that $U_n(\varphi) = \frac{1}{J}\sum_{j=1}^J \int_0^1 (\varphi_j \circ F_{n,j}^{-1} - F_{n,B}^{-1}(\varphi))^2$ and $U(\varphi) = \frac{1}{J}\sum_{j=1}^J \int_0^1 (\varphi_j \circ F_j^{-1} - F_B^{-1}(\varphi))^2$ with $F_{n,B}^{-1}(\varphi) = \frac{1}{J}\sum_{j=1}^J \varphi_j \circ F_{n,j}^{-1}$, $F_B^{-1}(\varphi) = \frac{1}{J}\sum_{j=1}^J \varphi_j \circ F_j^{-1}$. Now, (7.4.6) implies that $\sup_{\varphi_j \in \mathcal{G}_j} \int_0^1 (\varphi_j \circ F_j^{-1})^2 < \infty$. Similarly, assumption (7.4.5) implies

$$K_j := \sup_{\varphi_j \in \mathcal{G}_j, x \in (c_j,d_j)} |\varphi_j'(x)| < \infty.$$

Noting that $\int_0^1 (\varphi_j \circ F_{n,j}^{-1})^2 \leq 2\int_0^1 (\varphi_j \circ F_j^{-1})^2 + 2K_j^2 \int_0^1 (F_{n,j}^{-1} - F_j^{-1})^2$, we see that $\sup_{\varphi_j \in \mathcal{G}_j} \int_0^1 (\varphi_j \circ F_{n,j}^{-1})^2 < \infty$ a.s. and, with little additional effort, conclude that C_n has a.s. bounded trajectories. On the other hand, writing $d_{j,k}(\varphi) = \int_0^1 \varphi_j' \circ F_j^{-1} \frac{B_j}{f_j \circ F_j^{-1}} \varphi_k \circ F_k^{-1}$ we see that for $\varphi, \rho \in \mathcal{G}$

$$\begin{aligned}
|d_{j,k}(\varphi) - d_{j,k}(\rho)| &\leq \|\varphi_j' - \rho_j'\|_\infty \Big| \int_0^1 \frac{B_k}{f_k \circ F_k^{-1}} \varphi_k \circ F_k^{-1} \Big| \\
&+ \Big| \int_0^1 \rho_j' \circ F_j^{-1} \frac{B_k}{f_k \circ F_k^{-1}} (\varphi_k \circ F_k^{-1} - \rho_k \circ F_k^{-1}) \Big| \\
&\leq \|\varphi_j' - \rho_j'\|_\infty \sup_{\varphi_k \in \mathcal{G}_k} \Big| \int_0^1 \frac{B_k}{f_k \circ F_k^{-1}} \varphi_k \circ F_k^{-1} \Big| \\
&+ \sup_{(c_j,d_j)} |\rho_j'| \Big(\int_0^1 \Big| \tfrac{B_k}{f_k \circ F_k^{-1}} \Big|^q \Big)^{1/q} \Big(\int_0^1 |\varphi_k \circ F_k^{-1} - \rho_k \circ F_k^{-1}|^{p_0} \Big)^{1/p_0}
\end{aligned}$$

But using iii) of Lemma 7.7.1

$$\mathbb{E}\Big[\sup_{\varphi_k \in \mathcal{G}_k} \Big| \int_0^1 \frac{B_k}{f_k \circ F_k^{-1}} \varphi_k \circ F_k^{-1} \Big| \Big] \leq \int_0^1 \frac{\sqrt{t(1-t)}}{f_k(F_k^{-1}(t))} \sup_{\varphi_j \in \mathcal{G}_j} |\varphi_j(F_j^{-1}(t))| dt < \infty.$$

Hence, almost surely, $\sup_{\varphi \in \mathcal{G}} \Big| \int_0^1 \frac{B_j}{f_j \circ F_j^{-1}} \varphi_j \circ F_j^{-1} \Big| < \infty$. Furthermore, from assumption (7.4.3), we get that, a.s., $\int_0^1 \big(\frac{B_j}{f_j \circ F_j^{-1}}\big)^q < \infty$ and thus, for some a.s. finite random variable T,

$$|d_{j,k}(\varphi) - d_{j,k}(\rho)| \leq T \|\varphi - \rho\|_{\mathcal{G}}$$

for φ, $\rho \in \mathcal{G}$. From this conclude that the trajectories of C are a.s. bounded, uniformly continuous functions on $\mathcal{G}$, endowed with the norm $\|\cdot\|_{\mathcal{G}}$ introduced in (7.4.5). In particular, C is a tight random element in $L^\infty(\mathcal{G})$, see, e.g., p. 39-41 in Van der Vaart and Wellner.

From this point we pay attention to the quantile processes, namely,

$$\rho_{n,j}(t) = \sqrt{n} f_j(F_j^{-1}(t))(F_{n,j}^{-1}(t) - F_j^{-1}(t)), \quad 0 < t < 1, \quad j = 1, \ldots, J.$$

A trivial adaptation of Theorem 2.1, p. 381 in Csörgö and Horváth [1993] shows that, under (7.4.2), there exist, on a rich enough probability space, independent versions of $\rho_{n,j}$ and independent families of Brownian bridges $\{B_{n,j}\}_{n=1}\infty$, $j = 1, \ldots, J$, satisfying

$$n^{1/2-\nu} \sup_{1/n \leq t \leq 1-1/n} \frac{|\rho_{n,j}(t) - B_{n,j}(t)|}{(t(1-t))^\nu} = \begin{cases} O_p(\log(n)) \text{ if } \nu = 0 \\ O_p(1) \text{ if } 0 < \nu \leq 1/2 \end{cases} \tag{7.7.3}$$

We work, without loss of generality, with these versions of $\rho_{n,j}$ and $B_{n,j}$. We show now that

$$\sup_{\varphi\in\mathcal{G}}\Big|C_n(\varphi)-\hat{C}_n(\varphi)\Big|\to 0 \text{ in probability} \tag{7.7.4}$$

with $\hat{C}_n(\varphi)=\frac{1}{J}\sum_{j=1}^J c_{n,j}(\varphi)$ and $c_{n,j}(\varphi)=2\int_0^1 \varphi_j'\circ F_j^{-1}(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi))\frac{B_{n,j}}{f_j\circ F_j^{-1}}$. To check this we note that some simple algebra yields $C_n(\varphi)=\frac{2}{J}\sum_{j=1}^J\tilde{c}_{n,j}+\frac{1}{J}\sum_{j=1}^J\tilde{r}_{n,j}$ with

$$\tilde{c}_{n,j}=\sqrt{n}\int_0^1(\varphi_j\circ F_{n,j}^{-1}-\varphi_j\circ F_j^{-1})(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi)),$$

$$\tilde{r}_{n,j}=\sqrt{n}\int_0^1[(\varphi_j\circ F_{n,j}^{-1}-\varphi_j\circ F_j^{-1})-(F_{n,B}^{-1}(\varphi)-F_B^{-1}(\varphi))]^2.$$

From the elementary inequality $(a_1+\cdots+a_J)^2\le Ja_1^2+\cdots+Ja_J^2$ we get that

$$\frac{1}{J}\sum_{j=1}^J\tilde{r}_{n,j}\le\frac{4\sqrt{n}}{J}\sum_{j=1}^J\int_0^1(\varphi_j\circ F_{n,j}^{-1}-\varphi_j\circ F_j^{-1})^2\le\frac{4\sqrt{n}}{J}\sum_{j=1}^J K_j\int_0^1(F_{n,j}^{-1}-F_j^{-1})^2,$$

with $K_j:=\sup_{\varphi_j\in\mathcal{G}_j,x\in(c_j,d_j)}|\varphi_j'(x)|<\infty$, as above. Now we can use (7.4.4) and argue as in the proof of Theorem 2 in Alvarez-Esteban et al. [2008] to conclude that $\sqrt{n}\int_0^1(F_{n,j}^{-1}-F_j^{-1})^2\to 0$ in probability and, as a consequence, that

$$\sup_{\varphi\in\mathcal{G}}\Big|C_n(\varphi)-\frac{1}{J}\sum_{j=1}^J\tilde{c}_{n,j}(\varphi)\Big|\to 0 \text{ in probability.} \tag{7.7.5}$$

On the other hand, the Cauchy-Schwarz's inequality shows that

$$\begin{aligned}n\Big(\int_0^{\frac{1}{n}}(\varphi_j\circ F_{n,j}^{-1}-\varphi_j\circ F_j^{-1})(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi))\Big)^2&\\ \le\ \sqrt{n}\int_0^{\frac{1}{n}}(\varphi_j\circ F_{n,j}^{-1}-\varphi_j\circ F_j^{-1})^2\sqrt{n}\int_0^{\frac{1}{n}}(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi))^2&\end{aligned}$$

and using i) and ii) of Lemma 7.7.1, the two factors converge to zero uniformly in φ. A similar argument works for the upper tail and allows to conclude that we can replace in (7.7.5) $\tilde{c}_{n,j}(\varphi)$ with $\tilde{\tilde{c}}_{n,j}(\varphi):=2\sqrt{n}\int_{\frac{1}{n}}^{1-\frac{1}{n}}(\varphi_j\circ F_{n,j}^{-1}-\varphi_j\circ F_j^{-1})(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi))$. Moreover,

$$\sup_{\varphi\in\mathcal{G}}\Big|\int_0^{\frac{1}{n}}\varphi_j'\circ F_j^{-1}\frac{B_{n,j}}{f_j\circ F_j^{-1}}(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi))\Big|\le K_j\int_0^{\frac{1}{n}}\Big|\frac{B_{n,j}}{f_j\circ F_j^{-1}}\Big|\sup_{\varphi\in\mathcal{G}}\big|(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi))\big|$$

and by iii) of Lemma 7.7.1 and Cauchy-Schwarz's inequality

$$\mathbb{E}\Big[\int_0^{\frac{1}{n}}\Big|\frac{B_{n,j}}{f_j\circ F_j^{-1}}\Big|\sup_{\varphi\in\mathcal{G}}\big|(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi))\big|\Big]\le\int_0^{\frac{1}{n}}\frac{\sqrt{t(1-t)}}{f_j(F_j^{-1}(t))}\sup_{\varphi\in\mathcal{G}}\big|\varphi_j(F_j^{-1}(t))-F_B^{-1}(\varphi)(t)\big|dt\to 0.$$

Hence, $\sup_{\varphi\in\mathcal{G}}\Big|\int_0^{\frac{1}{n}}\varphi_j'\circ F_j^{-1}\frac{B_{n,j}}{f_j\circ F_j^{-1}}(\varphi_j\circ F_j^{-1}-F_B^{-1}(\varphi))\Big|\to 0$ in probability and similarly for the right tail. Now, for every $t\in(0,1)$ we have

$$\varphi_j\circ F_{n,j}^{-1}(t)-\varphi_j\circ F_j^{-1}(t)=\varphi_j'(K_{n,\varphi_j}(t))(F_{n,j}^{-1}(t)-F_j^{-1}(t)) \tag{7.7.6}$$

for some $K_{n,\varphi_j}(t)$ between $F_{n,j}^{-1}(t)$ and $F^{-1}(t)$. Therefore, (recall (7.7.6)), to prove (7.7.4) it suffices to show that

$$\sup_{\varphi\in\mathcal{G}}\Big|\int_{\frac{1}{n}}^{1-\frac{1}{n}}\varphi_j'(F_j^{-1}(t))\frac{B_{n,j}(t)}{f_j(F_j^{-1}(t))}(\varphi_j(F_j^{-1}(t))-F_B^{-1}(\varphi)(t))dt \tag{7.7.7}$$
$$-\int_{\frac{1}{n}}^{1-\frac{1}{n}}\varphi_j'(K_{n,\varphi_j}(t))\frac{\rho_{n,j}(t)}{f_j(F_j^{-1}(t))}(\varphi_j(F_j^{-1}(t))-F_B^{-1}(\varphi)(t))dt\Big|\to 0$$

in probability. To check it we take $\nu\in(0,1/2)$ in (7.7.3) to get

$$\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{|\rho_{n,j}(t)-B_{n,j}(t)|}{f_j(F_j^{-1}(t))}\sup_{\varphi\in\mathcal{G}}\big|\varphi_j(F_j^{-1}(t))-F_B^{-1}(\varphi)(t)\big|dt$$
$$\leq n^{\nu-\frac{1}{2}}O_P(1)\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{(t(1-t))^\nu}{f_k(F_k^{-1}(t))}\sup_{\varphi\in\mathcal{G}}\big|\varphi_j(F_j^{-1}(t))-F_B^{-1}(\varphi)(t)\big|dt\to 0 \tag{7.7.8}$$

in probability (using dominated convergence and iii) of Lemma 7.7.1). We observe next that, for each $t\in(0,1)$, $\sup_{\varphi_j\in\mathcal{G}_j}|K_{n,\varphi_j}(t)-F_j^{-1}(t)|\to 0$ a.s., since $K_{n,\varphi_j}(t)$ lies between $F_{n,j}^{-1}(t)$ and $F_j^{-1}(t)$. Therefore, using (7.4.5) we see that $\sup_{\varphi_j\in\mathcal{G}_j}|\varphi_j'(K_{n,\varphi_j}(t))-\varphi_j'(F_j^{-1}(t)|\to 0$ a.s. while, on the other hand, $\sup_{\varphi_j\in\mathcal{G}_j}|\varphi_j'(K_{n,\varphi_j}(t))-\varphi_j'(F_j^{-1}(t))|\leq 2K_j$. But then, by dominated convergence we get that

$$\mathbb{E}\Big[\sup_{\varphi_j\in\mathcal{G}_j}|\varphi_j'(K_{n,\varphi_j}(t))-\varphi_j'(F_j^{-1}(t))|^2\Big]\to 0.$$

Since by iii) of Lemma 7.7.1 we have that $t\mapsto\frac{\sqrt{t(1-t)}}{f_j(F_j^{-1}(t))}\sup_{\varphi\in\mathcal{G}}|\varphi_j(F_j^{-1}(t))-F_B^{-1}(\varphi)(t)|$ is integrable we conclude that

$$\mathbb{E}\sup_{\varphi\in\mathcal{G}}\int_{\frac{1}{n}}^{1-\frac{1}{n}}|\varphi_j'(K_{n,\varphi_j}(t))-\varphi_j'(F_j^{-1}(t))|\frac{|B_{n,j}(t)|}{f_j(F_j^{-1}(t))}|\varphi_j(F_j^{-1}(t))-F_B^{-1}(\varphi)(t)|dt$$

tends to 0 as $n\to\infty$ and, consequently,

$$\sup_{\varphi\in\mathcal{G}}\int_{\frac{1}{n}}^{1-\frac{1}{n}}|\varphi_j'(K_{n,\varphi_j}(t))-\varphi_j'(F_j^{-1}(t))|\frac{|B_{n,j}(t)|}{f_j(F_j^{-1}(t))}|\varphi_j(F_j^{-1}(t))-F_B^{-1}(\varphi)(t)|dt$$

vanishes in probability. Combining this fact with (7.7.8) we prove (7.7.7) and, as a consequence, (7.7.4). Finally, observe that for all $n\geq 1$, C has the same law as $\hat{C}_n$. This completes the proof. □

Proof of Theorem 7.4.1. From Skohorod Theorem (see, e.g., Theorem 1.10.4 in Van der Vaart and Wellner) we know that there exists on some probability space versions of C_n and C for which convergence of C_n to C holds almost surely. From now on, we place us on this space and observe that

$$\sqrt{n}(A_n(\mathcal{G})-A(\mathcal{G}))\leq\sqrt{n}\inf_\Gamma U_n-\sqrt{n}\inf_\Gamma U=\inf_{\varphi\in\Gamma}C_n(\varphi). \tag{7.7.9}$$

On the other hand, if we consider the (a.s.) compact set $\Gamma_n=\{\varphi\in\mathcal{G}:U(\varphi)\leq\inf_{\mathcal{G}}U+\frac{2}{\sqrt{n}}\|C_n\|_\infty\}$, then, if $\varphi\notin\Gamma_n$, $U_n(\varphi)\geq\inf_{\mathcal{G}}U+\frac{1}{\sqrt{n}}\|C_n\|_\infty$, while if $\varphi\in\Gamma$, then, $U_n(\varphi)\leq$

$\inf_{\mathcal{G}} U + \frac{1}{\sqrt{n}} \|C_n\|_\infty$. Thus, necessarily, $\inf_{\mathcal{G}} U_n = \inf_{\Gamma_n} U_n = \inf_{\Gamma_n}(U_n - U + U) \geq \inf_{\Gamma_n}(U_n - U) + \inf_{\Gamma_n} U = \inf_{\Gamma_n}(U_n - U) + \inf_{\Gamma} U$. Together with (7.7.9) this entails

$$\inf_{\varphi\in\Gamma_n} C_n(\varphi) \leq \sqrt{n}(A_n(\mathcal{G}) - A(\mathcal{G})) \leq \inf_{\varphi\in\Gamma} C_n(\varphi) \tag{7.7.10}$$

Note that for the versions that we are considering $\|C_n - C\|_\infty \to 0$ a.s.. In particular, this implies that $\inf_\Gamma C_n \to \inf_\Gamma C$ a.s.. Hence, the proof will be complete if we show that a.s.

$$\inf_{\Gamma_n} C_n \to \inf_{\Gamma} C. \tag{7.7.11}$$

To check this last point, consider a sequence $\varphi_n \in \Gamma_n$ such that $C_n(\varphi_n) \leq \inf_{\Gamma_n} C_n + \frac{1}{n}$. By compactness of $\mathcal{G}$, taking subsequences if necessary, $\varphi_n \to \varphi_0$ for some $\varphi_0 \in \mathcal{G}$. Continuity of U yields $U(\varphi_n) \to U(\varphi_0)$ and as a consequence, that $U(\varphi_0) \leq \inf_{\mathcal{G}} U$, that is, $\varphi_0 \in \Gamma$ a.s.. Furthermore,

$$\left|C_n(\varphi_n) - C(\varphi_0)\right| \leq \|C_n - C\|_\infty + |C(\varphi_n) - C(\varphi_0)| \to 0.$$

This shows that

$$\liminf \inf_{\Gamma_n} C_n \geq C(\varphi_0) \geqslant \inf_{\Gamma} C \tag{7.7.12}$$

and yields (7.7.11). This completes the proof.

□

Proof of Corollary 7.4.2. In Theorem 7.3.3, take $\mu'_j = \mu_{n,j}$. Then, writing $\mathcal{L}^*$ for the conditional law given the $X_{i,j}$, the result of Theorem 7.3.3 reads

$$\mathcal{W}_2^2(\mathcal{L}((A_{m_n}(\mathcal{G}))^{1/2}), \mathcal{L}^*((A^*_{m_n}(\mathcal{G}))^{1/2})) \leq L^2 \frac{1}{J} \sum_{j=1}^{J} \mathcal{W}_2^2(\mu_j, \mu_{n,j}),$$

with $L = \sup_{\varphi\in\mathcal{G}} \left\|\varphi'_j\right\|_\infty < \infty$. Since $\mathcal{W}_r(\mathcal{L}(aX + b), \mathcal{L}(aY + b)) = a\mathcal{W}_r(\mathcal{L}(X), \mathcal{L}(Y))$ for $a > 0, b \in \mathbb{R}$, the last bound gives

$$\begin{aligned}\mathcal{W}_2^2(\mathcal{L}\big(\sqrt{m_n}\big((A_{m_n}(\mathcal{G}))^{1/2} - (A(\mathcal{G}))^{1/2}\big)\big), \mathcal{L}^*\big(\sqrt{m_n}\big((A^*_{m_n}(\mathcal{G}))^{1/2} - (A(\mathcal{G}))^{1/2}\big)\big)) \\ \leq \quad L^2 \frac{m_n}{\sqrt{n}} \frac{1}{J} \sum_{j=1}^{J} \sqrt{n}\mathcal{W}_2^2(\mu_j, \mu_{n,j}).\end{aligned}$$

As noted in the proof of Theorem 7.4.1, the assumptions imply that $\sqrt{n}\mathcal{W}_2^2(\mu_j, \mu_{n,j})$ vanishes in probability. Also, Theorem 7.4.1 and the delta method yield that

$$\sqrt{m_n}\big((A_{m_n}(\mathcal{G}))^{1/2} - (A(\mathcal{G}))^{1/2}\big) \rightharpoonup \frac{1}{2(A(\mathcal{G}))^{1/2}}\gamma,$$

with γ the limiting law there, which, combined to the above bound, shows that

$$\sqrt{m_n}\big((A^*_{m_n}(\mathcal{G}))^{1/2} - (A(\mathcal{G}))^{1/2}\big) \rightharpoonup \frac{1}{2(A(\mathcal{G}))^{1/2}}\gamma$$

in probability. A further use of the delta method yields

$$\sqrt{m_n}\big(A^*_{m_n}(\mathcal{G}) - A(\mathcal{G})\big) \rightharpoonup \gamma$$

in probability. The result follows now from Lemma 1 in Janssen and Pauls [2003]. □

Proof of Theorem 7.5.1. We assume for simplicity that $p = 1$. The general case follows with straightforward changes. Let us observe that

$$U_n(\theta) = \frac{1}{J}\sum_{j=1}\int_0^1 (\psi_j(\theta_j, G_{n,j}^{-1}) - \tfrac{1}{J}\textstyle\sum_{k=1}^J \psi_k(\theta_k, G_{n,k}^{-1}))^2,$$

with $G_{n,j}$ the empirical d.f. on the $\varepsilon_{i,j}$'s (which are i.i.d. G). A similar expression, replacing $G_{n,j}$ with G is valid for $U(\theta)$. Then (7.5.6) implies that $\sup_\theta |U_n(\theta) - U(\theta)| \to 0$, from which (recall (7.5.3) it follows that $\hat{\theta}_n \to \theta^*$ in probability. Note that the second part in Assumption (7.5.6) is a technical assumption that ensures that, when considering a Taylor expansion in the integral of $U_n(\theta)$, the remainder term in $\psi_j'(\lambda, H_{n,j}^{-1}) - \psi_j'(\lambda, G_j^{-1})$ for any $H_{n,j}^{-1}$ lying between $G_{n,j}^{-1}$ and G_j^{-1} (obtained through a Taylor expansion) goes uniformly to zero.

From (7.5.4) we have that U_n is a C^2 function whose derivatives can be computed by differentiation under the integral sign. This implies that

$$\begin{aligned} D_j U_n(\theta) &= \frac{2}{J}\int_0^1 D\psi_j(\theta_j, G_{n,j}^{-1})(\psi_j(\theta_j, G_{n,j}^{-1}) - \tfrac{1}{J}\textstyle\sum_{k=1}^J \psi_k(\theta_k, G_{n,k}^{-1})), \\ D_{p,q}U_n(\theta) &= -\frac{2}{J^2}\int_0^1 D\psi_p(\theta_p, G_{n,p}^{-1}) D\psi_q(\theta_q, G_{n,q}^{-1}), \quad p \neq q \end{aligned} \tag{7.7.13}$$

and

$$\begin{aligned} D_{p,p}U_n(\theta) &= \frac{2}{J}\int_0^1 D^2\psi_p(\theta_p, G_{n,p}^{-1})(\psi_j(\theta_j, G_{n,j}^{-1}) - \tfrac{1}{J}\textstyle\sum_{k=1}^J \psi_k(\theta_k, G_{n,k}^{-1})) \\ &\quad + \frac{2(J-1)}{J^2}\int_0^1 (D\psi_p(\theta_p, G_{n,p}^{-1}))^2. \end{aligned}$$

Using also (7.5.5) we obtain similar expressions for the derivatives of $U(\theta)$, replacing everywhere $G_{n,j}^{-1}$ with G^{-1}. We write $DU_n(\theta) = (D_j U_n(\theta))_{1\le j\le J}$, $DU(\theta) = (D_j U(\theta))_{1\le j\le J}$ for the gradients and $\Sigma_n(\theta) = [D_{p,q}U_n(\theta)]_{1\le p,q\le J}$, $\Sigma(\theta) = [D_{p,q}U(\theta)]_{1\le p,q\le J}$for the Hessians of U_n and U. Note that $\Sigma^* = \Sigma(\theta^*)$ is assumed to be invertible.

We write now $\rho_{n,j}$ for the quantile process based on the $\varepsilon_{i,j}$'s. Observe that (7.5.7) ensures that we can assume, without loss of generality, that there exist independent Brownian bridges, $B_{n,j}$, satisfying (7.7.3). Now, recalling that $\psi_j(\theta_j^*, x) = x$ we see that

$$\sqrt{n}D_j U_n(\theta^*) = \frac{2}{J}\int_0^1 D\psi_j(\theta_j^*, G_{n,j}^{-1}(t))\frac{\rho_{n,j}(t) - \frac{1}{J}\sum_{k=1}^J \rho_{n,k}(t)}{g(G^{-1}(t))}dt. \tag{7.7.14}$$

Now, using (7.5.5) and arguing as in the proof of Theorem 7.4.1 we conclude that

$$\left|\int_0^1 D\psi_j(\theta_j^*, G_{n,j}^{-1}(t))\frac{\rho_{n,k}(t)}{g\left(G^{-1}(t)\right)}dt - \int_0^1 D\psi_j(\theta_j^*, G^{-1}(t))\frac{B_{n,k}(t)}{g\left(G^{-1}(t)\right)}dt\right| \to 0$$

in probability and, consequently,

$$\left|\sqrt{n}D_j U_n(\theta^*) - \frac{2}{J}\int_0^1 D\psi_j(\theta_j^*, G^{-1}(t))\frac{B_{n,j}(t) - \frac{1}{J}\sum_{k=1}^J B_{n,k}(t)}{g\left(G^{-1}(t)\right)}dt\right| \to 0 \tag{7.7.15}$$

in probability.

A further Taylor expansion of $D_j U_n$ around θ^* shows that for some $\tilde{\theta}_j^n$ between $\hat{\theta}_n$ and θ^* we have

$$D_j U_n(\hat{\theta}_n) = D_j U_n(\theta^*) + (D_{1j}U_n(\tilde{\theta}_j^n), \ldots, D_{Jj}^2 U_n(\tilde{\theta}_j^n)) \cdot (\hat{\theta}_n - \theta^*)$$

and because $\hat{\theta}_n$ is a zero of DU_n, we obtain

$$-D_j U_n(\theta^*) = (D_{1j}U_n(\tilde{\theta}_j^n), \ldots, D_{Jj}U_n(\tilde{\theta}_j^n)) \cdot (\hat{\theta}_n - \theta^*).$$

Writing $\tilde{\Sigma}_n$ for the $J \times J$ matrix whose J-th row equals $(D_{1j}U_n(\tilde{\theta}_j^n), \ldots, D_{Jj}U_n(\tilde{\theta}_j^n))$, $j = 1, \ldots, J$, we can rewrite the last expansion as

$$-\sqrt{n}DU_n(\theta^*) = \tilde{\Sigma}_n \sqrt{n}(\hat{\theta}_n - \theta^*). \tag{7.7.16}$$

Now, recalling (7.7.13), assumptions (7.5.4) and (7.5.5) yield that $\tilde{\Sigma}_n \to \Sigma^* = \Sigma(\theta^*)$ in probability. As a consequence, (7.7.16) and (7.7.15) together with Slutsky's Theorem complete the proof of the second claim.

Finally, for the proof of the last claim, since $DU_n(\hat{\theta}_n) = 0$, a Taylor expansion around $\hat{\theta}_n$ shows that

$$nU_n(\theta^*) - nU_n(\hat{\theta}_n) = \frac{1}{2}(\sqrt{n}(\hat{\theta}_n - \theta^*))' \Sigma(\tilde{\theta}_n)(\sqrt{n}(\hat{\theta}_n - \theta^*)) \tag{7.7.17}$$

for some $\tilde{\theta}_n$ between $\hat{\theta}_n$ and θ^*. Arguing as above we see that $\Sigma(\tilde{\theta}_n) \to \Sigma^*$ in probability. Hence, to complete the proof if suffices to show that

$$nU_n(\theta^*) - \frac{1}{J}\sum_{j=1}^{k} \int_0^1 \frac{\left(B_{n,j}(t) - \frac{1}{J}\sum_{k=1}^{J} B_{n,k}(t)\right)^2}{g(G^{-1}(t))^2} dt \to 0$$

in probability. Since

$$nU_n(\theta^*) = \frac{1}{J}\sum_{j=1}^{k} \int_0^1 \frac{\left(\rho_{n,j}(t) - \frac{1}{J}\sum_{k=1}^{J} \rho_{n,k}(t)\right)^2}{g(G^{-1}(t))^2} dt,$$

this amounts to proving that

$$\int_0^1 \frac{\left(\rho_{n,j}(t) - B_{n,j}(t)\right)^2}{g(G^{-1}(t))^2} dt \to 0$$

in probability. Taking $\nu \in (0, \frac{1}{2})$ in (7.7.3) we see that

$$\int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{\left(\rho_{n,j}(t) - B_{n,j}(t))\right)^2}{g(G^{-1}(t))^2} dt \le O_P(1) \frac{1}{n^{1-2\nu}} \int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{(t(1-t))^{2\nu}}{g(G^{-1}(t))^2} \to 0,$$

using condition (7.5.8) and dominated convergence. From (7.5.8) we also see that

$$\int_{1-\frac{1}{n}}^{1} \frac{B_{n,j}(t)^2}{g(G^{-1}(t))^2} dt \to 0 \text{ in probability.}$$

Condition (7.5.8) implies also that $\int_{1-\frac{1}{n}}^{1} \frac{\rho_{n,j}(t)^2}{g(G^{-1}(t))^2} dt \to 0$ in probability, see Samworth and Johnson [2004]. Similar considerations apply to the left tail and complete the proof.

□

Proof of Corollary 7.5.2. Writing $\mathcal{L}^*$ for the conditional law given the $X_{i,j}$'s, we see from Theorem 7.3.3 that

$$\mathcal{W}_2^2(\mathcal{L}(\sqrt{m_n}(A_{m_n}(\Theta))^{1/2}), \mathcal{L}^*(\sqrt{m_n}(A^*_{m_n}(\Theta))^{1/2}) \le L\frac{m_n}{n}\frac{1}{J}\sum_{j=1}^{J} n\mathcal{W}_2^2(\mu, \tilde{\mu}_{n,j}),$$

where $L = \sup_{\lambda,x,j} \psi_j'(\lambda, x)$, μ denotes the law of the errors, $\varepsilon_{i,j}$, and $\tilde{\mu}_{n,j}$ the empirical d.f. on $\varepsilon_{1,j}, \ldots, \varepsilon_{n,j}$. Note that $L < \infty$ by (7.5.6), while $n\mathcal{W}_2^2(\mu, \tilde{\mu}_{n,j}) = O_P(1)$ as in the proof of Theorem 7.5.1. Hence, we conclude that

$$m_n A^*_{m_n}(\Theta) \rightharpoonup \frac{1}{J} \sum_{j=1}^{J} \int_0^1 \Big(\frac{\tilde{B}_j}{g \circ G^{-1}} \Big)^2 - \frac{1}{2} Y^T \Sigma^{-1} Y$$

in probability. The conclusion now follows from Lemma 1 in Janssen and Pauls [2003].

□

If centering were necessary and we had (7.5.11) rather than the limit in Theorem 7.5.1 we could adapt the last argument as follows. If A and B are positive random variables then $E|A-B| \leq E(A^{1/2}-B^{1/2})^2+2(EAE(A^{1/2}-B^{1/2})^2)^{1/2}$. We can apply this bound to (an optimal coupling of) $m_n A_{m_n}(\Theta)$ and $m_n A^*_{m_n}(\Theta)$. Now if the errors have a log-concave distribution then $nE\mathcal{W}_2^2(\mu, \tilde{\mu}_{n,j}) = O(\log n)$, see Corollary 6.12 in Bobkov and Ledoux [2014] and we conclude that

$$\mathcal{W}_1(\mathcal{L}(m_n A_{m_n}(\Theta) - c_{m_n}), \mathcal{L}^*(m_n A^*_{m_n}(\Theta) - c_{m_n})) = \mathcal{W}_1(\mathcal{L}(m_n A_{m_n}(\Theta)), \mathcal{L}^*(m_n A^*_{m_n}(\Theta)))$$

vanishes in probability if $m_n = O(n^\rho)$ for some $\rho \in (0, 1)$. As a consequence,

$$m_n A^*_{m_n}(\Theta) - c_{m_n} \rightharpoonup \frac{1}{J} \sum_{j=1}^{J} \int_0^1 \frac{\tilde{B}_j^2 - E\tilde{B}_j^2}{(g \circ G^{-1})^2} - \frac{1}{2} Y^T \Sigma^{-1} Y$$

in probability.

7.7.3 Tables

Table 7.1 – Simulations under H_0

J	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
	50	0,144	0,079	0,038	0,046	0,041	0,03
	100	0,148	0,067	0,07	0,05	0,04	0,033
	200	0,129	0,085	0,068	0,043	0,037	0,044
2	500	0,138	0,089	0,05	0,048	0,035	0,036
	1000	0,127	0,086	0,063	0,055	0,039	0,032
	2000	0,129	0,104	0,071	0,048	0,043	0,038
	5000	0,039	0,042	0,041	0,049	0,043	0,055
	50	0,295	0,194	0,115	0,078	0,054	0,034
	100	0,273	0,163	0,089	0,053	0,034	0,039
	200	0,238	0,15	0,077	0,054	0,047	0,031
3	500	0,226	0,122	0,07	0,057	0,042	0,029
	1000	0,217	0,107	0,092	0,069	0,042	0,035
	2000	0,221	0,128	0,077	0,053	0,043	0,035
	5000	0,205	0,145	0,082	0,06	0,025	0,047
	50	0,659	0,428	0,281	0,129	0,111	0,081
	100	0,583	0,337	0,192	0,104	0,083	0,053
	200	0,538	0,281	0,159	0,081	0,078	0,029
5	500	0,449	0,267	0,138	0,063	0,056	0,04
	1000	0,415	0,238	0,129	0,064	0,051	0,037
	2000	0,354	0,212	0,115	0,06	0,053	0,032
	5000	0,322	0,203	0,108	0,057	0,061	0,039
	50	0,996	0,971	0,873	0,702	0,553	0,456
	100	0,994	0,902	0,708	0,433	0,33	0,226
	200	0,958	0,802	0,521	0,247	0,184	0,119
10	500	0,914	0,663	0,388	0,149	0,093	0,063
	1000	0,864	0,532	0,286	0,119	0,084	0,046
	2000	0,813	0,473	0,239	0,103	0,063	0,051
	5000	0,756	0,449	0,217	0,088	0,061	0,041

Table 7.2 – Power of the test for $\gamma \stackrel{d}{=} \varepsilon\,(1)$

J	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
2	50	0,961	0,919	0,897	0,864	0,829	0, 838
	100	1	0,998	0,998	0,995	0,994	0,993
	200	1	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
3	50	0,987	0,971	0,97	0,953	0,939	0,91
	100	1	1	0,999	1	0,999	0,999
	200	1	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
5	50	1	0,996	0,988	0,976	0,971	0,955
	100	1	1	1	1	1	1
	200	1	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
10	50	1	1	1	1	0,996	0,985
	100	1	1	1	1	1	1
	200	1	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1

Table 7.3 – Power of the test $\gamma \stackrel{d}{=} Laplace\,(0,1)$

J	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
2	50	0,426	0,33	0,3	0,241	0,223	0,163
	100	0,658	0,534	0,468	0,365	0,361	0,3
	200	0,855	0,824	0,751	0,665	0,613	0,602
	500	0,998	0,998	0,993	0,982	0,965	0,962
	1000	1	1	1	1	0,999	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
3	50	0,657	0,533	0,422	0,331	0,282	0,223
	100	0,831	0,708	0,586	0,514	0,461	0,377
	200	0,946	0,915	0,841	0,778	0,709	0,661
	500	1	0,998	0,997	0,994	0,989	0,977
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
5	50	0,895	0,741	0,633	0,471	0,394	0,333
	100	0,936	0,874	0,728	0,623	0,519	0,443
	200	0,994	0,947	0,903	0,847	0,786	0,696
	500	1	1	1	0,996	0,992	0,985
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
10	50	1	0,997	0,97	0,875	0,79	0,703
	100	0,997	0,985	0,949	0,854	0,765	0,643
	200	1	0,996	0,968	0,924	0,859	0,789
	500	1	1	1	0,996	0,996	0,975
	1000	1	1	1	1	1	0,999
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1

Table 7.4 – Power of the test $\gamma \stackrel{d}{=} T(3)$

I	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
	50	0,566	0,445	0,429	0,352	0,321	0,307
	100	0,775	0,704	0,647	0,576	0,503	0,454
	200	0,942	0,927	0,882	0,833	0,771	0,697
2	500	1	0,997	0,995	0,991	0,989	0,957
	1000	1	1	1	1	1	0,986
	2000	1	1	1	1	1	0,999
	5000	1	1	1	1	1	0,997
	50	0,745	0,653	0,546	0,46	0,402	0,349
	100	0,881	0,821	0,738	0,65	0,592	0,563
	200	0,98	0,958	0,928	0,891	0,873	0,794
3	500	1	1	0,999	0,997	0,997	0,978
	1000	1	1	1	1	1	0,995
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
	50	0,91	0,813	0,682	0,593	0,525	0,45
	100	0,972	0,909	0,822	0,751	0,686	0,621
	200	0,995	0,984	0,967	0,915	0,887	0,836
5	500	1	1	1	0,999	0,999	0,995
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
	50	1	0,997	0,953	0,894	0,827	0,758
	100	0,999	0,993	0,969	0,907	0,862	0,79
	200	1	0,998	0,995	0,961	0,941	0,903
10	500	1	1	1	1	0,998	0,988
	1000	1	1	1	1	1	0,998
	2000	1	1	1	1	1	0,999
	5000	1	1	1	1	1	1

Table 7.5 – Power of the test $\gamma \stackrel{d}{=} T(4)$

I	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0.95}$	$m_n = n$
	50	0,398	0,353	0,292	0,207	0,182	0,183
	100	0,623	0,52	0,429	0,341	0,29	0,228
	200	0,826	0,717	0,65	0,589	0,526	0,41
2	500	0,989	0,978	0,954	0,928	0,878	0,787
	1000	1	1	0,999	1	0,984	0,955
	2000	1	1	1	1	1	0,985
	5000	1	1	1	1	1	0,993
	50	0,634	0,495	0,4	0,295	0,263	0,222
	100	0,756	0,666	0,56	0,465	0,399	0,336
	200	0,914	0,859	0,778	0,663	0,602	0,521
3	500	0,998	0,989	0,985	0,972	0,928	0,868
	1000	1	1	1	1	0,999	0,963
	2000	1	1	1	1	1	0,989
	5000	1	1	1	1	1	1
	50	0,851	0,709	0,583	0,426	0,359	0,316
	100	0,919	0,825	0,668	0,546	0,493	0,316
	200	0,959	0,908	0,842	0,738	0,684	0,578
5	500	1	0,997	0,994	0,973	0,934	0,888
	1000	1	1	1	1	0,999	0,968
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	0,999
	50	1	0,986	0,941	0,813	0,774	0,653
	100	1	0,988	0,925	0,806	0,738	0,606
	200	1	0,991	0,948	0,854	0,813	0,679
10	500	1	1	0,998	0,985	0,954	0,886
	1000	1	1	1	1	0,997	0,949
	2000	1	1	1	1	1	0,974
	5000	1	1	1	1	1	0,995

Concluding remarks and future work

The generalization of the use of machine learning algorithms in the everyday life and the professional world has been accompanied by concerns about the ethical issues that may arise from the adoption of these technologies. Even more, the entire population is becoming increasingly aware of its serious implications. Therefore, the notion of fairness in machine learning has received a growing interest among the research community over the last years, resulting in a great push for the emergence of multidisciplinary approaches for assessing and removing the presence of bias in algorithms. We believe this is crucial in order to guarantee a fair treatment for every subgroup of population, which will contribute to reduce the growing distrust of machine learning systems in the society. This thesis aims at studying the recent established area of fair learning through an optimal transport based approach. We summarize in the following the main contributions and mention some of the possible lines of future work.

In the first part, we have motivated the fairness problem by presenting a case-study of the use of machine learning techniques for the prediction of the real and well-known benchmark *Adult Income* dataset. In particular, we have provided some comprehensive results from the analysis of the fairness criterion *statistical parity* measured through the *disparate impact* index, for which we have proposed an ad-hoc construction of confidence intervals. This metric quantifies the difference between the behaviour of a classification rule applied for two subgroups of the population, the minority and the majority. Fairness is achieved when the algorithm behaves in the same way for both groups, resulting in the sensitive variable not playing a significant role in the prediction. Importantly, we have noticed that trying to make fair machine learning models may be a particularly challenging task, especially when the training observations contain bias. In such cases, standard regulations that promote either the removal of the sensitive variable or the use of testing techniques appeared as irrelevant when dealing with fairness of machine learning algorithms. This content is available online in Besse et al. [2020] and currently submitted for publication. We have also provided a companion notebook at `https://github.com/XAI-ANITI/StoryOfBias/blob/master/StoryOfBias.ipynb` for reproducibility purposes.

Then we have presented a review of mathematical models designed to handle the issue of bias in machine learning in a general setting. We have proposed a probabilistic approach to characterize *perfect fairness* in terms of the independence between the sensitive attribute and the outcome of the algorithm, or conditional independence when the true value of the target is available in the learning data. Within both frameworks, we have defined and then computed the so-called price for fairness to quantify the real impact of fairness constraint on the behavior of a machine learning algorithm. We have provided some novel contributions in the analysis of this price in regression and classification. When perfect fairness requires to pay a too high price, resulting in poor generalization errors with respect to the unfair case, it is natural not to impose this strict condition but rather weaken the fairness constraint. A review of the methods for imposing a level of fairness has been presented, with a classification into *pre-*, *in-* or *post-processing* methods, depending on the time of application of the fairness conditions. We have

noticed that, while a substantial part of the models in the first and last families are based on optimal transport, methods in the *in-processing* group, which includes the majority of the contributions in the literature, can be seen as fair risk minimization problems.

Our study provides a better understanding of fair learning, yet many cases remain open to further research to obtain a full theoretical framework. We have pointed out that we did not consider in this study many new interesting points of view on fairness that deserve a specific study, including a causal approach for fairness [Loftus et al., 2018] or using counter-examples [Kusner et al., 2017, Black et al., 2020].

In the particular case of classification, we have recasted the links between fairness and predictability in terms of probability metrics. We have analyzed a repairing methodology based on mapping conditional distributions to the Wasserstein barycenter, which is included in the first category mentioned above. As a main contribution, we have justified such approach providing an upper bound for the price for fairness of the transportation towards the barycenter. Finally, we have proposed a *random repair* which yields a tradeoff between minimal information loss and a certain amount of fairness. This content was presented at the *International Conference of Machine Learning (Los Angeles, june 2019)* and it is therefore published in the book of Proceedings of Machine Learning Research as Gordaliza et al. [2019].

The second part of the thesis has been devoted to the asymptotic theory of the empirical transportation cost. First, we have provided a Central Limit Theorem for the Monge-Kantorovich distance $\mathcal{W}_p(P_n, Q_m)$ between two empirical distributions with different sizes n and m for observations on $\mathbb{R}$ and general cost $p \geq 1$. In the case $p > 1$ our assumptions are sharp in terms of moments and smoothness. We have also proved results dealing with the choice of centering constants. With important implications for statistical inference, we have obtained a consistent estimate of the asymptotic variance which enables to build two sample tests and confidence intervals to certify the similarity between two distributions. These have then been used to assess a new criterion of dataset fairness in classification. These contributions correspond to the publication del Barrio et al. [2019b]. Additionally, we have provided a moderate deviation principle for the empirical transportation cost in general dimension.

Finally, Wasserstein barycenters and variance-like criterion in terms of the Wasserstein distance are used in many problems to analyze the homogeneity of collections of distributions and structural relationships between the observations. In del Barrio et al. [2019a] we have proposed the estimation of the quantiles of the empirical process of the Wasserstein's variation using a bootstrap procedure. Then we have used these results for statistical inference on a distribution registration model for general deformation functions. The tests are based on the variance of the distributions with respect to their Wasserstein's barycenters, for which we have proved central limit theorems, including bootstrap versions. Although a detailed study on the application of these results to fair learning remains for future work, a rough idea has been outlined in the introduction of the thesis. Precisely, we have noticed that the problem of repairing the data could be addressed through a deformation model.

The future work offered by this line of research is as broad as it is unpredictable, given the dizzying evolution that artificial intelligence, particularly machine learning and data science, is currently undergoing along with society's misgivings and concerns.

Admittedly, there is still a long way to go in the deepening in fair learning and its mathematical basis. As mentioned above, some methodologies including the causal approach or the so-called counterfactual fairness have not been addressed in this thesis, but deserve further attention. Additionally, it is worth considering the extension of our optimal transport based approach to fair learning into other methodological contexts such as graphical [Baer et al., 2019, Gilbert, 2019] or econometric models.

Besides this essential theoretical deepening in the mathematical models for fair learning, the future of this promising area of machine learning must be aware of the fact that the main motivation for its development actually relies on the wide range of real problems in which fairness plays an important role. Let us mention at least two areas in which the growing of fair learning is more than necessary as well as promising. First of all, the industrial application of fair learning is a clear emerging area particularly characterized, on one hand, by being almost free of ethical issues and, on the other hand, by the clear economic return that can be expected in fields such as image processing for computer vision, statistical quality control and so on. Secondly, but not less important at all, we must mention health applications, especially relevant nowadays due to the pandemic situation that we are living, caused by COVID-19 disease. It is clear that the use of contact tracking apps or the governments likely issuing biological passports push the importance of fairness in machine learning algorithms as a priority to deal with.

Finally, as machine learning is an emerging area in rapid and continuous development, it will be necessary to analyze the connections between fair learning and other areas of machine learning, such as transfer learning or domain adaptation. In particular, optimal transport techniques seem to be apropriate to deal with these interesting problems.

Bibliography

P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

A. Agarwal, M. Dudík, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. *arXiv preprint arXiv:1905.12843*, 2019.

M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

M. Agulló-Antolín, J. A. Cuesta-Albertos, H. Lescornel, and J.-M. Loubes. A parametric registration model for warped distributions with wasserstein's distance. *Journal of Multivariate Analysis*, 135:117–130, 2015.

M. Ajtai, J. Komlós, and G. Tusnády. On optimal matchings. *Combinatorica*, 4:259–264, 1984.

J. Ali, M. B. Zafar, A. Singla, and K. P. Gummadi. Loss-aversively fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 211–218, 2019.

S. Allassonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

P. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. I. H. Poincaré-Pr.*, 47: 358–375, 2011.

P. C. Alvarez-Esteban, E. Del Barrio, J. A. Cuesta-Albertos, and C. Matran. Trimmed comparison of distributions. *Journal of the American Statistical Association*, 103(482):697–704, 2008.

Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *Journal of the American Statistical Association*, 86(414):376–387, 1991.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

B. R. Baer, D. E. Gilbert, and M. T. Wells. Fairness criteria through the lens of directed acyclic graphical models. *arXiv preprint arXiv:1906.11333*, 2019.

M. Ballu, Q. Berthet, and F. Bach. Stochastic optimization for regularized wasserstein estimators. *arXiv preprint arXiv:2002.08695*, 2020.

S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

A. Barron. Uniformly powerful goodness of fit tests. *Ann. Statist.*, 17:107–124, 1989.

F. Barthe and N. O'Connell. Matchings and the variance of lipschitz functions. *ESAIM: Probability and Statistics*, 13:400–408, 2009.

Y. Bechavod and K. Ligett. Penalizing Unfairness in Binary Classification. *ArXiv e-prints*, June 2017.

Y. Bechavod and K. Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov): 2399–2434, 2006.

R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017a.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017b.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

P. Berthet, J.-C. Fort, and T. Klein. A central limit theorem for wasserstein type distances between two different laws. *arXiv preprint arXiv:1710.09763*, 2017.

P. Besse, C. Castets-Renard, A. Garivier, and J.-M. Loubes. L'ia du quotidien peut elle être éthique? 2018a.

P. Besse, E. del Barrio, P. Gordaliza, and J.-M. Loubes. Confidence intervals for testing disparate impact in fair learning. *arXiv preprint arXiv:1807.06362*, 2018b.

P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Risser. A survey of bias in machine learning through the prism of statistical parity for the adult data set. *arXiv preprint arXiv:2003.14263v2*, 2020.

A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

D. Biddle. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

J. Bigot. Statistical data analysis in the wasserstein space. *arXiv preprint arXiv:1907.08417*, 2019.

J. Bigot and T. Klein. Characterization of barycenters in the wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

E. Black, S. Yeom, and M. Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics and kantorovich transport distances. *preprint*, 2014.

E. Boissard, T. Le Gouic, and J.-M. Loubes. Distributions template estimate with Wasserstein metrics. *Bernoulli*, 21:740–759, 2015.

B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514–560, 2005.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized preprocessing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

S. Chiappa, R. Jiang, T. Stepleton, A. Pacchiano, H. Jiang, and J. Aslanides. A general approach to fairness with optimal transport. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

F. Chierichetti, S. Kumar, R.and Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037, 2017.

L. Chizat, G. Peyré, Schmitzer, B., and F. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.*, 2018.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*, pages 12739–12750, 2019.

O. Collier and A. S. Dalalyan. Curve registration by nonparametric goodness-of-fit testing. *Journal of Statistical Planning and Inference*, 162:20–42, 2015.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. *arXiv preprint arXiv:1807.00028*, 2018.

J. S. Cramer. The origins of logistic regression. 2002.

M. Csörgö and L. Horváth. *Weighted approximations in probability and statistics*. J. Wiley & Sons, 1993.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

C. Czado and A. Munk. Assessing the similarity of distributions-finite sample performance of the empirical mallows distance. *Journal of Statistical Computation and Simulation*, 60(4): 319–346, 1998.

Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

E. del Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019.

E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez. Tests of goodness-of-fit based on the l_2-Wasserstein distance. *Ann. Statist.*, pages 1230–1239, 1999a.

E. del Barrio, E. Giné, and C. Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.*, pages 1009–1071, 1999b.

E. del Barrio, E. Giné, F. Utzet, et al. Asymptotics for l2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11:131–189, 2005.

E. del Barrio, P. Gordaliza, H. Lescornel, and J.-M. Loubes. Central limit theorem and bootstrap procedure for wasserstein's variations with an application to structural relationships between distributions. *Journal of Multivariate Analysis*, 169:341–362, 2019a.

E. del Barrio, P. Gordaliza, and J.-M. Loubes. A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8(4):817–849, 2019b.

E. del Barrio, P. Gordaliza, and J.-M. Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98406-2. doi: 10.1007/978-1-4612-5320-4.

W. Dieterich, C. Mendoza, and T. Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 2016.

V. Dobrić and J. E. Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8:97–118, 1995.

N. A. Doherty, A. V. Kartasheva, and R. D. Phillips. Information effect of entry into credit ratings market: The case of insurers' ratings. *Journal of Financial Economics*, 106(2):308–330, 2012.

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical Risk Minimization under Fairness Constraints. *ArXiv e-prints*, Feb. 2018.

M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

R. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

R. M. Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

J. Dunkelau and M. Leuschel. Fairness-aware machine learning.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

H. Edwards and A. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations*, 2015.

H. Edwards and A. Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

M. Feldman. *Computational Fairness: Preventing Machine-Learned Discrimination*. PhD thesis, 2015.

S. A. Feldman, M.and Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

J. Fish, B.and Kun and A. D. Lelkes. Fair boosting: a case study. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Citeseer, 2015.

A. W. Flores, K. Bechtel, and C. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Rel.*, 162:707–738, 2015.

G. Freitag and A. Munk. On hadamard differentiability in k-sample semiparametric models—with applications to the assessment of structural relationships. *Journal of multivariate analysis*, 94(1):123–158, 2005.

G. Freitag, C. Czado, and A. Munk. A nonparametric test for similarity of marginals-with applications to the assessment of population bioequivalence. *J. Stat. Plan. Infer.*, 137:697–711, 2007.

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

K. Fukuchi, T. Kamishima, and J. Sakuma. Prediction with model-based neutrality. *IEICE TRANSACTIONS on Information and Systems*, 98(8):1503–1516, 2015.

S. Gallón, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical biosciences*, 242(2):129–142, 2013.

F. Gamboa, J.-M. Loubes, and E. Maza. Semi-parametric estimation of shifts. *Electronic Journal of Statistics*, 1:616–640, 2007.

A. Ganesh and N. O'Connell. Large and moderate deviations for matching problems and empirical discrepancies. *Markov Process. Related Fields*, 13(1):85–98, 2007.

A. Gano. Disparate impact and mortgage lending: A beginner's guide. *U. Colo. L. Rev.*, 88: 1109, 2017.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018.

A. Ghassami, S. Khodadadian, and N. Kiyavash. Fairness in supervised learning: An information theoretic approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 176–180. IEEE, 2018.

D. E. Gilbert. Luck, fairness and bayesian tensor completion. 2019.

G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

P. Gordaliza, E. del Barrio, F. Gamboa, and J.-M. Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365, 2019.

N. Gozlan and C. Léonard. A large deviation approach to some transportation cost inequalities. *Probability Theory and Related Fields*, 139(1-2):235–283, 2007.

N. Gozlan and C. Léonard. Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*, 2010.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129, 2005.

P. Hacker and E. Wiedemann. A continuous framework for fairness. *CoRR*, abs/1712.07924, 2017.

S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer, and F. Giannotti. Injecting discrimination and privacy awareness into pattern discovery. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 360–369. IEEE, 2012.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

U. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948, 2018.

M. Hurley and J. Adebayo. Credit scoring in the era of big data. *Yale JL & Tech.*, 18:148, 2016.

J.-C. Hütter and P. Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

A. Janssen and T. Pauls. How do bootstrap and permutation tests work? *The Annals of statistics*, 31(3):768–806, 2003.

R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. *arXiv preprint arXiv:1907.12059*, 2019.

J. E. Johndrow and K. Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6. Citeseer, 2010.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874, Dec 2010. doi: 10.1109/ICDM.2010.50.

F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 1991.

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

M. P. Kim, A. Ghorbani, and J. Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

A. Komiyama, J.and Takeda, J. Honda, and H. Shimao. Nonconvex optimization for regression with fairness constraints. In *International conference on machine learning*, pages 2737–2746, 2018.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

P. Lahoti, K. P. Gummadi, and G. Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

T. Le Gouic and J.-M. Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

T. Le Gouic and J.-M. Loubes. Computing the price for fairness in a regression framework. *arXiv preprint arXiv:2005.11720*, 2020.

M. Ledoux. Sur les déviations modérées des sommes de variables aléatoires vectorielles indépendantes de même loi. In *Annales de l'IHP Probabilités et statistiques*, volume 28, pages 267–280, 1992.

Y. L'Horty, M. Bunel, S. Mbaye, P. Petit, and L. du Parquet. Discriminations dans l'accès à la banque et à l'assurance.

Z. Li, A. Perez-Suay, G. Camps-Valls, and D. Sejdinovic. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *arXiv preprint arXiv:1911.04322*, 2019.

J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *ArXiv e-prints*, Oct. 2016.

D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018a.

D. Madras, T. Pitassi, and R. Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pages 6147–6157, 2018b.

J. Mary, C. Calauzenes, and N. El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391, 2019.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18: 1269–1283, 1990.

P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4543–4553. Curran Associates, Inc., 2019.

A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

M. Mercat-Bruns. *Discrimination at Work*. University of California Press, 2016.

Q. Mérigot, A. Delalande, and F. Chazal. Quantitative stability of optimal transport maps and linearization of the 2-wasserstein space. *arXiv preprint arXiv:1910.05954*, 2019.

T. M. Mitchell et al. Machine learning, 1997.

G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein training of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 3718–3726, 2016.

S. B. Morris and R. E. Lobsenz. Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology*, 53(1):89–111. doi: 10.1111/j.1744-6570.2000.tb00195.x.

A. Munk and C. Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60:223–241, 1998.

R. Nabi and I. Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

H. Narasimhan. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654, 2018.

J. Niles-Weed and P. Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

A. Noriega-Campero, M. A. Bakker, B. Garcia-Bulle, and A. Pentland. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 77–83, 2019.

V. Noroozi, S. Bahaadini, S. Sheikhi, N. Mojab, and P. S. Yu. Leveraging semi-supervised learning for fairness using neural networks. *arXiv preprint arXiv:1912.13230*, 2019.

L. Oneto and S. Chiappa. Fairness in machine learning. In *Recent Trends in Learning From Data*, pages 155–196. Springer, 2020.

L. Oneto, M. Donini, A. Elders, and M. Pontil. Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 227–237, 2019.

F. Pasquale. *The black box society*. Harvard University Press, 2015.

D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM international conference on data mining*, pages 581–592. SIAM, 2009.

D. Pedreschi, S. Ruggieri, and F. Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 126–131. ACM, 2012.

D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 339–355. Springer, 2017.

G. Peyré, M. Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

N. Quadrianto and V. Sharmanska. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems*, pages 677–688, 2017.

S. T. Rachev. The Monge-Kantorovich problem on mass transfer and its applications in stochastics. *Teor. Veroyatnost. i Primenen.*, 29:625–653, 1984. ISSN 0040-361X.

S. T. Rachev. *Probability metrics and the stability of stochastic models*, volume 269. John Wiley & Son Ltd, 1991.

M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 469–481, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372828. URL https://doi.org/10.1145/3351095.3372828.

J. Ramsay and B. Silverman. Functional data analysis. *Springer series in statistics*, pages 10–18, 2005.

I. Redko, A. Habrard, and M. Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 737–753. Springer, 2017.

A. Rezaei, R. Fathony, O. Memarrast, and B. Ziebart. Fairness for robust log loss classification.

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

P. A. Riach and J. Rich. Field experiments of discrimination in the market place. *The economic journal*, 112(483):F480–F518, 2002.

M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

P. Rigollet and J. Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235, 2018.

T. Rippl, A. Munk, and A. Sturm. Limit laws of the empirical wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109, 2016.

L. Risser, Q. Vincenot, N. Couellan, and J.-M. Loubes. Using wasserstein-2 regularization to ensure fair decisions with neural-network classifiers. *arXiv preprint arXiv:1908.05783*, 2019.

R. T. Rockafellar and R. J.-B. Wets. Variational analysis. grundlehren series (fundamental principles of mathematical sciences), vol. 317, 1998.

A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.*, 29:582–638, 2014a.

A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582?638, 2014b. doi: 10.1017/S0269888913000039.

R. Rothmann, J. Krieger-Lamina, and W. Peissl. Credit scoring in Österreich, 07 2014.

R. Samworth and O. Johnson. Convergence of the empirical process in mallows distance, with an application to bootstrap performance. *arXiv preprint math/0406603*, 2004.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

M. Serrurier, J.-M. Loubes, and E. Pauwels. Fairness with wasserstein adversarial networks. Technical report, working paper or preprint, 2019.

P. W. Shor. *Random planar matching and bin packing*. PhD thesis, Massachusetts Institute of Technology, Department of Mathematics, 1985.

G. Shorack. *Probability for Statisticians*. Springer, 2000.

Y. R. Shrestha and Y. Yang. Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms*, 12(9):199, 2019.

R. B. Siegel. Race-conscious but race-neutral: The constitutionality of disparate impact in the roberts court. *Ala. L. Rev.*, 66:653, 2014.

C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 80:219–238, 2018.

J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, 2018.

T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, and M. B. Weller, A.and Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, 2018.

D. Steinberg, A. Reid, S. O'Callaghan, F. Lattimore, L. McCalman, and T. Caetano. Fast fair regression via efficient approximations of mutual information. *arXiv preprint arXiv:2002.06200*, 2020.

C. D. Sutton. Classification and regression trees, bagging, and boosting. *Handbook of statistics*, 24:303–329, 2005.

M. Talagrand and J. Yukich. The integrability of the square exponential transportation cost. *Ann. App. Probab.*, pages 1100–1111, 1993.

C. Tameling, M. Sommerfeld, and A. Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *arXiv preprint arXiv:1707.00973*, 2017.

Z. Tan, S. Yeom, M. Fredrikson, and A. Talwalkar. Learning fair representations for kernel models. *arXiv preprint arXiv:1906.11813*, 2019.

G. Torrisi. Asymptotic analysis of the optimal cost in some transportation problems with random locations. *Stochastic Processes and their Applications*, 122(1):305–333, 2012.

A. Van der Vaart and J. Wellner. Weak convergence and empirical processes. 1996.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

S. Verma and J. Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Soc., 2003. ISBN 9780821833124. URL `https://books.google.es/books?id=GqRXYFxe0l0C`.

C. Villani. *Optimal transport: old and new*. Springer Verlag, 2009.

B. Winrow and C. Schieber. The disparity between disparate treatment and disparate impact: An analysis of the ricci case. *Academy of Legal, Ethical and Regulatory Issues*, page 27, 2009.

B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

L. Wu. Large deviations, moderate deviations and lil for empirical processes. *The Annals of Probability*, 22(1):17–27, 1994.

G. Yona and G. Rothblum. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688, 2018.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017a.

M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017b.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

M. Zehlike, P. Hacker, and E. Wiedemann. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery*, 34(1):163–200, 2020.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

X. Zhang and M. Liu. Fairness in learning-based sequential decision algorithms: A survey. *arXiv preprint arXiv:2001.04861*, 2020.

I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.